THE EVOLUTION OF BIG DATA AND ITS BUSINESS APPLICATIONS

Marwah Ahmed Halwani

Dissertation Prepared for the Degree of

DOCTOR OF PHILOSOPHY

UNIVERSITY OF NORTH TEXAS

May 2018

APPROVED:

Victor R. Prybutok, Major Professor and Dean of the Toulouse Graduate School
Daniel Peak, Committee Co-Chair
Brian O'Connor, Committee Member
Nicholas Evangelopoulos, Committee Member
Suliman Hawamdeh, Chair of the Department of Information Science
Kinshuk, Dean of the College of Information

Halwani, Marwah Ahmed. *The Evolution of Big Data and Its Business Applications*. Doctor of Philosophy (Information Science), May 2018, 113 pp., 12 tables, 10 figures, references, 203 titles.

The arrival of the Big Data era has become a major topic of discussion in many sectors because of the premises of big data utilizations and its impact on decision-making. It is an interdisciplinary issue that has captured the attention of scholars and created new research opportunities in information science, business, heath care, and many others fields. The problem is the Big Data is not well defined, so that there exists confusion in IT what jobs and skill sets are required in big data area. The problem stems from the newness of the Big Data profession. Because many aspects of the area are unknown, organizations do not yet possess the IT, human, and business resources necessary to cope with and benefit from big data. These organizations include health care, enterprise, logistics, universities, weather forecasting, oil companies, e-business, recruiting agencies etc., and are challenged to deal with high volume, high variety, and high velocity big data to facilitate better decision- making. This research proposes a new way to look at Big Data and Big Data analysis. It helps and meets the theoretical and methodological foundations of Big Data and addresses an increasing demand for more powerful Big Data analysis from the academic researches prospective. Essay 1 provides a strategic overview of the untapped potential of social media Big Data in the business world and describes its challenges and opportunities for aspiring business organizations. It also aims to offer fresh recommendations on how companies can exploit social media data analysis to make better business decisions—decisions that embrace the relevant social qualities of its customers and their related ecosystem. The goal of this research is to provide insights for businesses to make better, more informed decisions based on effective social media data analysis. Essay 2 provides a better understanding

of the influence of social media during the 2016 American presidential election and develops a model to examine individuals' attitudes toward participating in social media (SM) discussions that might influence their decision in choosing between the two presidential election candidates, Donald Trump and Hilary Clinton. The goal of this research is to provide a theoretical foundation that supports the influence of social media on individual's decisions. Essay 3 defines the major job descriptions for careers in the new Big Data profession. It to describe the Big Data professional profile as reflected by the demand side, and explains the differences and commonalities between company-posted job requirements for data analytics, business analytics, and data scientists jobs. The main aim for this work is to clarify of the skill requirements for Big Data professionals for the joint benefit of the job market where they will be employed and of academia, where such professionals will be prepared in data science programs, to aid in the entire process of preparing and recruiting for Big Data positions.

ACKNOWLEDGMENTS

In the name of God, the Most Gracious, the Most Merciful.

As Abu Huraira reported: The Prophet, peace, and blessings be upon him, said, “He has not thanked Allah who has not thanked people” (Sunan Abī Dāwūd, 4811). I would like to express my deepest thanks to my advisor and chairman of the committee, Dr. Victor Prybotuk, and Co-chair, Dr. Daniel Peak. Their guidance, encouragement, and support have been the largest factor behind the success of the dissertation research. I also would like to thank Dr. Nick Evangelopoulos, for giving me the opportunity to work with him, learn, and grow not only as researcher but also as a person. I sincerely appreciate Dr. Brian O’Conner for his support. Working with all of them has been very enjoyable, challenging, and motivating.

For my husband, Yaser Banoun. I am grateful for his support, his patience, and the sacrifices he made to help me accomplish my dream, and to do the best throughout this journey. I am blessed for my kids’, Sarah, Hasan, and Sidrah, understanding and patience. Their inspiration provides me with the strongest support.

For my parents, Ahmed Halwani, and Hayat Alsulimany, there are not enough words to describe how thankful I am to both of them. Thank you for always being there for my family and me. I am who I am because of your strength, dedication, and prayers. Also, special thanks to my sisters, Dr. Manal, Mai, and Maha, and to my brother, Mohammed, for all their encouragement.

My sincere gratitude to all my friends, especially Aleka Myre and Terri Ann Metzler, for helping me, as they are my family. I also want to take the opportunity to thank King Abdulaziz University for honoring me with full funded scholarship. I look forward to be able to give back many years of hard work.

Above all, I would like to thank the Almighty Allah for making all this possible.

TABLE OF CONTENTS

Page

ACKNOWLEDGMENTS iii

LIST OF TABLES vii

LIST OF FIGURES viii

CHAPTER 1. INTRODUCTION 1

1.1 Background 1

1.1.1 The Definition of Information 1

1.1.2 The Origins of Information Science 3

1.1.3 Information Science, Information Technology, Information Systems (MIS), and Business 5

1.1.4 The Differences between Information Technology and Information Systems 5

1.1.5 Information Systems (MIS) and Information Science 7

1.1.6 The Impact of Information Science on Business Practices involving Big Data Phenomena and Big Data Analysis 9

1.1.7 The Evolution of Big Data 10

1.1.8 The Evolution of the Professions 17

1.2 Problem Statement 18

1.3 Research Question 18

1.4 Purpose and Contribution 19

1.5 Research Design 20

1.6 Organization of the Dissertation 21

CHAPTER 2. LITERATURE REVIEW 22

2.1 The Definition of Big Data 22

2.2 History of Big Data 24

2.2.1 Decision Support Systems (DSSs) 24

2.2.2 Decision Support Systems (DSSs) 24

2.2.3 The 1970s Period: Decision Support Systems (DSSs) 25

2.2.4 The 1980s Period: Enterprise/Executive Information Systems 26

2.2.5 The 1990s Period: Business Intelligence 27

2.2.6 The 2000s Period: Analytics 29
2.3 Big Data Sources: Social Media Sites 32
2.4 The Reasons for Studying Big Data from Three Different Approaches 33

CHAPTER 3. METHODOLOGY 36
3.1 Essay 1 Value of theoretical foundation based on literature review 36
3.2 Essay 2 Survey Research 36
3.2.1 Perceived Usefulness and Perceived Ease of Use 38
3.2.2 Social Media Norm (SN) and Social Media Community Identification (CI) 38
3.2.3 Individual Attitudes 39
3.2.4 Perceived Social Influence 39
3.3 Essay 3 Latent Semantic Analysis 41
3.3.1 Data Collection 41
3.3.2 Text Analytics 41

CHAPTER 4. RESULTS AND DISCUSSION 44
4.1 Essay 1: Using Social Media Data Contributes to Better Business Decisions 44
4.1.1 Introduction 44
4.1.2 Literature Review 46
4.2 Essay 2 Social Media's Influence on Millennials: A Case Study of the 2016 American Presidential Election 62
4.2.1 Introduction 62
4.2.2 Literature Review/Theoretical Background 66
4.2.3 Research Model and Hypotheses 70
4.2.4 Methodology 72
4.2.5 Data Analysis and Results 74
4.2.6 Discussion 77
4.2.7 Conclusions 78
4.3 Essay 3 Big Data Professions 78
4.3.1 Introduction 79
4.3.2 Big Data Professions 81
4.3.3 Big Data Professions Responsibilities 85
4.3.4 Methodology 87

4.3.5 Results 90
4.3.6 Discussion 94
4.3.7 Conclusion 96

CHAPTER 5. CONCLUSION 97

REFERENCES 99

LIST OF TABLES

Page

Table 2.1: Evolution of Big Data terminology (Sharda et al., 2017) 24

Table 3.1: Respondent demographics 40

Table 4.1: Analogy of the donut strategy for several SM platforms 49

Table 4.2: Respondent demographics 74

Table 4.3: Measurement model summary. 75

Table 4.4: Candidates * individuals' attitudes categories cross tabulation 76

Table 4.5: Chi-square tests 77

Table 4.6: Big Data profession responsibilities 86

Table 4.7: Topic extraction 90

Table 4.8: The contingency table 92

Table 4.9: Job categories on the principal component space 93

Table 4.10: Qualification topics on the principal component space 93

LIST OF FIGURES

Page

Figure 1.1: Ackoff's pyramid 2

Figure 1.2: ISc, computer science, IS and management (Dbmesser, 2011) 6

Figure 1.3: The research concept 10

Figure 1.4: Big Data process life cycle 13

Figure 3.1: Conceptual framework 37

Figure 3.2: Scree plot of eigenvalues 43

Figure 4.1: Conceptual framework 70

Figure 4.2: PLS structural equation path results. Path significant at $P < 0.05$ level 76

Figure 4.3: Scree plot of eigenvalues 89

Figure 4.4: Correspondence map showing a two-dimensional projection of the 15-by-9 matrix 94

CHAPTER 1

INTRODUCTION

1.1 Background

1.1.1 The Definition of Information

There is no universalTable definition of information; however, according to Targowski, information is best defined from the perspective of a discipline. He applies the model of the Semantic Ladder to define the notions of data, information, knowledge, and wisdom in terms of cognition units. Targowski also describes the connection between information and wisdom.

The quantitative perspective on information defines information as "a view as the successful selection of signs or words that form a given list, rejecting all "semantic meaning" as a subjective factor" (p. 55). Alternatively, the definition of information from the qualitative perspective is in terms of a set of eight qualitative attributes: relevance, timeliness, exclusiveness, format, accessibility, accuracy, verifiability, and price. From the cognitive perspective on information - business perspective, the definition of information is "A comparative unit of cognition that defines a change between the previous and present state of natural, artificial, or semantic systems" (Targowski, 2005, p. 61). The cognitive perspective provides this definition of information after clarifying the definition of the four conative units following Ackoff's pyramid or the knowledge pyramid shown in Figure 1.1. Ackoff's pyramid presents the relationships among data, information, knowledge, and wisdom (Ackoff, 1989). The concept of data is that of raw facts, which have no significant value by themselves, and can have different formats. Information is data that are organized because the data was processed into a form that is meaningful and has value. Knowledge is collected information that creates awareness. All the concepts lead

to the definition of wisdom, which is a vision and depth of understanding acquired through extensive knowledge and experience (Ackoff, 1989; Bellinger, Castro, & Mills, 2004).

Figure 1.1: Ackoff's pyramid

Moreover, from the information science perspective, Shannon & Weaver 1949, developed the information theory that studies the transmission, processing, extraction, and utilization of information. The theory includes basic element of communication (Shannon, & Weaver, 1949): 1) An information is a source of a message, 2) A person who transmits the message to create a signal which can be sent through a channel is called transmitter, 3) The medium that the signal carries the information sent over it is called a channel, 4) A person who transforms the signal back into the message intended for delivery is called a receiver, and 5) A person or a machine, for whom or which the message is intended is called a destination (Shannon, & Weaver, 1949).

In addition, Belkin and Robertson (1976) proposed two premises regarding the fundamental concepts of information science. First, the discipline of information science is considered to be problem-oriented due to the effective transferring of information from the generators to users. Second, the common concept of information is extended to include the idea of the change of information structure. The authors defined information as "the structure of any text which is capable of changing the image-structure of a recipient," and they used the definition of

information and text to discuss the fundamental concepts of information science (Belkin & Robertson, 1976, p. 201).

A definition particularly relevant to my research interest is that of information fusion. Information fusion is defined as "an Information Process dealing with the association, correlation, and combination of data and information from single and multiple sensors or sources to achieve refined estimates of parameters, characteristics, events, and behaviors for observed entities in an observed field of view. It is sometimes implemented as a Fully Automatic process or as a Human-Aiding process for Analysis and/or Decision Support." (Llinas, 2001. p.1; Boström, & Andler, 2007. p. 3). In addition, McKinney, et al. defined information with four views: token which is information and data are both tokens manipulated by processes, syntax which is information is the measurable relationship among tokens that reduces entropy, representation: information is meaning, and adaptation where is the subjectivist assumptions are introduced to explain how information is created by a system (e.g., person, organization) (2010).

1.1.2 The Origins of Information Science

Borko (1968) discussed information science and defined the term information scientist to help clarify the nature of the information science field and work within the field. Information science is "an interdisciplinary science that investigates the properties and behavior of information, the forces that govern the flow and use of information, and the techniques, both manual and mechanical, of processing information for optimal storage, retrieval, and dissemination" (p.5). Also, the author stated the main goal of information science is to spread knowledge everywhere. Historically, transmitting knowledge is a social responsibility, and the actual background of information science is this social responsibility (Belkin & Robertson, 1986). A study by Bates

(1999) clarified the fundamental keys of the information science standard and highlighted the role of information science as a meta-science by conducting a study to develop a theory around the documentary products of other disciplines and activities. Also, this article discusses how professions focus on the representation and organization of information rather than on knowing information or describing the fundamental skills of information representation in other professions and intellectual disciplines (Bates, 1999).

Commonly, information science "has both a pure science component, which reviews into the subject without regard to its application and an applied science component, which develops services and products" (Borko, 1968). To exemplify the relationship between the two sciences, Borko (1968) illustrates that there is a connection between research and application and between theory and practice, which means that every improvement that has been developed through research also needs to be tested through its application. Applied information science is concerned with the form of the information rather than the content. On the other hand, content is the domain of interest of pure science (Bates, 1999). From the social science perspective, information science is related to people's knowledge and culture (Buckland, 2012).

Information science consists of two major traditions: (1) documentation, which covers archives, bibliography, documentation, librarianship, records, and management, and (2) the computational tradition, which covers algorithmic, logical, mathematical, and mechanical techniques (Buckland, 1999). Impressively, Buckland (1999) has outlined the landscape of information science as a discipline. He learned that the information science landscape is very complex, and the reason for this complexity is the knowledge and the way in which it is important for all fields. The following are aspects of the landscape of information science as Buckland (1999) described them:

(1) Information science study (computing, mathematics, language, ethnography, semiotics, economics, and law).

(2) Information science specialties (geographical information systems, socio-economic datasets, and websites).

(3) Information science application contexts (restaurants, libraries, travel agencies, medical clinics, and universities).

(4) Information science ideological (economic, political, cultural) situations.

1.1.3 Information Science, Information Technology, Information Systems (MIS), and Business

Generally, as mentioned above, information science is an interdisciplinary field; it focuses on the analysis, collection, classification, manipulation, storage, retrieval, movement, dissemination, and protection of information. This means that information science is the umbrella discipline for many other disciplines such as computer science, management sciences, social science, business information systems, etc. Figure 1.2 shows how the focuses of my dissertation topic narrow down from information science, to information systems, business, and my particular research interest, the Big Data application.

1.1.4 The Differences between Information Technology and Information Systems

Information systems (IS) and information technology (IT) are similar in many ways, but they are also different. Researchers have struggled and differed in defining IS and IT and showing the relationship between them. Several researchers assume that IT is a part of the IS environment, as illustrated in Figure 1.2. They define IT as "technologies developed in the computational science-based disciplines, including computer science, information science, and electrical and electronic engineering" (Dbmeseer, 2011; Wang, 2010). Others define IS in these terms:

> An information system is a social system which has embedded in it information technology The extent to which information technology plays a part is increasing rapidly.

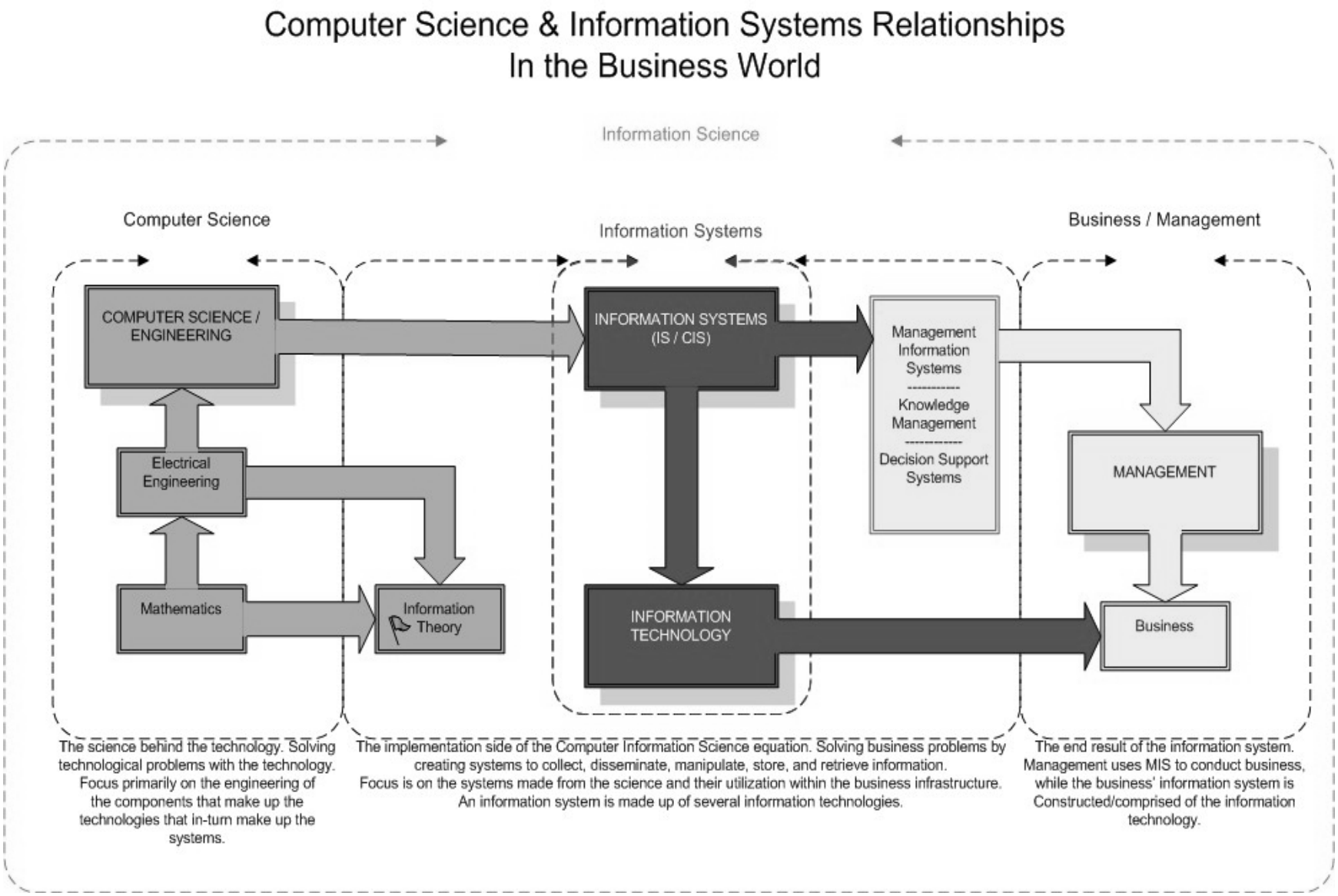

Figure 1.2: ISc, computer science, IS and management (Dbmesser, 2011)

> But this does not prevent the overall system from being a social system, and it is not possible to design a robust, effective information system, incorporating significant amounts of the technology without treating it as a social system. (Land 1985, p. 215, Magalhães 1999, p. 6)

We agree with the concept that IS is the process part of the IT environment. According to the Information Technology Association of America, information technology is defined as "the study, design, development, application, implementation, support or management of computer-based information systems" (Balust, & Macario, 2009). On the other hand, The United Kingdom Academy for Information Systems (UKAIS) defines information systems as "the means by which organizations and people, utilizing information technologies, gather, process store, use and disseminate information" (Monarch, 2000, p. 3; Ellis, Allen, & Wilson, 1999). Also, Huber et al. defined IS as "an organized collection of people, information, business processes, and information technology designed to transform inputs into outputs, in order to achieve a goal" (2007, p. 392).

1.1.5 Information Systems (MIS) and Information Science

As mentioned previously, the main goal of the information science (ISc) discipline is to provide information appropriately to help people improve their knowledge and be more educated. Borko stated that "Information science as a discipline has as its goal to provide a body of information that will lead to improvement in the various institutions and procedures dedicated to the accumulation and transmission of knowledge. There are in existence a number of such institutions and related media" (1968, p. 3). For example, information science manages the contents of education to make them useful for teachers and to transform the knowledge for learners. As Borko (1968) illustrated the connection between pure science and applied science by giving examples of research and application, Bates also framed the practical and the theoretical level of information science using the examples of a physician and actors representing information. The

recording of information and the creating of databases represent forms of information (1999). This illustrates the function of information science with information systems. According to the Institute for Information Scientists (IIS), Information Science is "broad concepts and theories of information systems and information and communication technologies insofar as they apply to the principles and practices of information management" (Webber, 2003, p. 313).

Nowadays, information systems have become the major practical means to achieve the goal of the information science discipline. Alter describes information systems as "a system in which human participants and/or machines perform work (processes and activities) using information, technology, and other resources to produce informational products and/or services for internal or external customers" (2008, p. 6). A study by Ellis (1999) investigated the relationships between the fields of information systems and information science and found there is an overlap between the two disciplines. Following Ellis, researchers conducted a meta-analysis of titles and abstracts in journals in information science, information systems, and medical. The author found that the overlap between information systems and information science is artificial. The ISc discipline is concerned about the content and how recorded information and knowledge can be organized and accessed to develop applicable information services. While IS is concerned about formal organizational relationships and how information can be used to develop effectively and some further objectives from the data (Monarch, 2000; Beeson, & Chelin, 2006). The artificial concept can be seen in many other fields such as healthcare, education, business, and science.

Monarch also presents an example of other fields to illustrate the relationship between ISc and IS such as medical informatics (MI) and information retrieval (IR) and explains how these fields share several research interests. For instance, the information system leximap and the information science leximap show clear common subject matter in Decision Supporting

Systems (DSS), specifically in medical informatics healthcare organizations (Monarch, 2000). Healthcare organizations use clinical decision support systems and other systems. Using effective systems in the medical field is important to improve clinical practice and meet medical information needs. Among the recent research on information in practice, seventy studies indicate that the clinical medical practice is significantly improved, by 68%, when decision support systems are used (Kawamoto, Houlihan, Balas, & Lobach, 2005). We focus on the business management information system in this research, figure 4 illistrates the research concepts.

1.1.6 The Impact of Information Science on Business Practices involving Big Data Phenomena and Big Data Analysis

Business is a global discipline and touches every aspect of individual's lives. Business processes, personal economics, finance, investment, accounting, management, information systems, and data management are all areas that impact individuals daily. Information Sciences, including information systems, help business professionals how to use the latest technology to meet business needs. Using information is a way to explore problem-solving in the area of business issues to improve business decisions, or business operations and analytics. A study by Sevrani (2011) addressed the impact of using IS/IT in business processes to achieve success in various markets by studying the use of IS/IT in Albania, a developing country; it concluded that Albanian businesses should use more IS/IT to improve their business processes and decision.

Information scientists recognize the problem of information management in business organizations such as information overload or data overload. Information overload has no generally agreed upon definition yet; the term may refer to having much relevant information, or a large amount of irrelevant data from different sources due to the advancements of technology

(Edmunds & Morrise, 2000; Butcher, 1998). It is clear that information overload is becoming established as a concept relevant to Big Data nowadays and the field of data sciences.

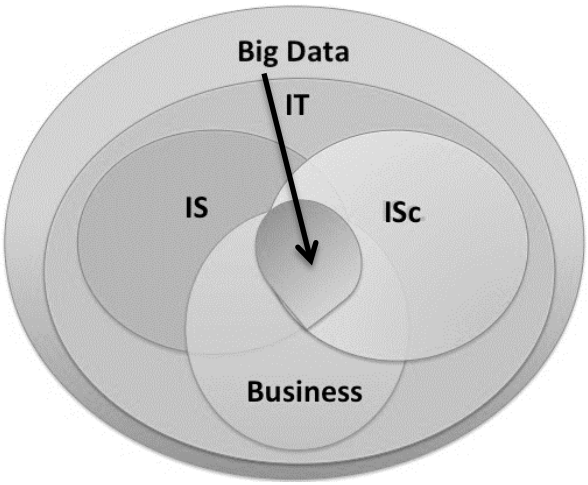

Figure 1.3: The research concept

1.1.7 The Evolution of Big Data

The charateristics of data differentiate and define Big Data. These characteristics include volume, variety, and velocityand are known as the 3 V's and the 4 V's when value is included (Kim, Trimi, & Chung, 2014). Several researchers have examined these characteristics in terms of ther attributes and challenges. The term volume relates to data size in terabytes (TB), petabytes (PB), or zettabytes (ZB), and variety refers to the types of data. Velocity refers to how often the data are produced, and the fourth V is the value the information provides (Zaslavsky, Perera, & Georgakopoulos, 2013).

An IBM report (2013) shows that company's view of evolution of Big Data from 1910 to 2013 using infographics and animations. In summary they show that in 1919, the U.S. Department of commerce employees used more than 200 million IBM-supplied punched cards and processing equipment capturing massive amounts of data to perform the US Agricultural Census. In 1923, the

Los Angeles Police Department used IBM tabulating equipment and collected and analyzed data to identify criminal methods, boosting crime-solving efforts, and provided insights to improve the quality an efficiency of the police department (IBM, 2013).

In 1934, IBM developed a flagship tabulation product, the 405 Electric Accounting Machine. That machine processed 150 80-character punched cards a minute and printed alphanumeric results at a rate of 802 characters a minute. In 1956, RAMAC was introduced by IBM, Random Access Method of Accounting and Control. It is the first magnetic hard disk for data storage and could store about 2000 bits of data per square inch at a cost of $10,000 per megabyte. In 1969, IBM technology guided the *Apollo* mission to the moon, which was designed by NASA. The *Saturn* instrument unit was a computer nerve center for the launch vehicle that processed data and controlled the *Saturn* rocket until *Apollo* safely headed to the moon. In 1970, relational databases were introduced by IBM, they stored information in the computer that was arranged within tables allowing easier access and data management (IBM, 2013).

In 1988, IBM provided six 3090 Model 300E mainframes with vector facilities that allowed the National Institutes of Health to increase incremental CPU capacity by 35% annually and these improved data handling and access methods allowed breakthroughs in the health industry. In 2002, the “eDiamond” project was built by Oxford University, IBM, and the U.K. government. That project developed a computing Grid for early screening and diagnosis of breast cancer, and provided medical professionals with more information about treating the disease. In 2005, IBM presented a new automotive industry business solution that allows automakers and fleet owners to collect and analyze large amounts of data about their vehicles to improve the identification of trends, better manage warranty coverage, and help the manufacturers adhere to government regulations (IBM, 2013).

In 2009, the SmartBay project was developed via a collaboration between IBM and the Marine Institute of Ireland. SmartBay was created to collect environmental data and involved the use of analytics to provide insights on pollution levels and environmental conditions. In 2011, IBM launched the Watson computing system Watson understands natural language and can analyze data and find correlations. In 2013, IBM collaborated with more than 1000 global universities to help developing a Big Data and analytics curriculum. As part of their work with the universities, a set of job qualifications was also developed (IBM, 2013). The evolution of Big Data also shows that the term "Big Data" is not simply about huge amounts of data but that it also encompasses the related processes of gathering, storing and analyzing that data. The following is an explanation of Big Data process lifecycle.

1.1.7.1 Big Data Capture and Store

According to experts, Big Data such as click-stream data are captured from the data sources and distributed across multiple nodes, which are usually in the form of a grid using MapReduce and its open source implementation Hadoop (Ramakrishnan & Ghoshal, 2014). Each section of the grid processes a subset of data in parallel (Dyché, 2012). Hadoop is the Apache open-source software framework for distributed storage and processing of large data sets on computer clusters built from commodity hardware. Hadoop consists of two components, the storage component, called Hadoop Distributed File System (HDFS), and the processing component, called MapReduce (Hadoop, 2009). MapReduce is a parallel processing model for processing and generating large data sets (Dean & Ghemawat, 2008).

Previously, most companies collected structured data from their daily transactions and

stored them in a database, which were two parts of the data process. Today, with the variety and velocity of large volumes of data, there exists a huge challenge associated with capturing and using those data. Not only is capturing data a potential challenge for Big Data processes depending on the data types, sources, and fields, but also it is a challenge to clean the data and make sure that that data are accurate and on time. Several studies have taken up the issue of capturing Big Data. Researchers have examined the integration of artificial intelligence (AI) and Big Data. They have learned that AI has been used in several different ways to facilitate the capturing and structuring of Big Data, and there are some challenging issues associated with emerging AI and Big Data (O'Leary, 2013).

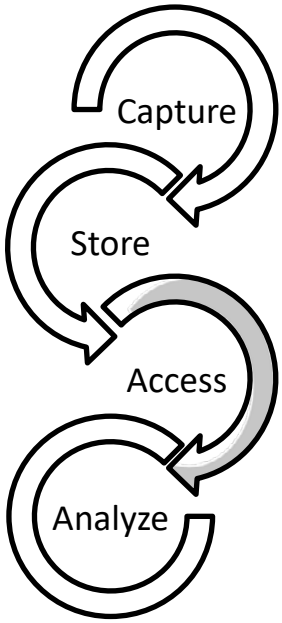

Figure 1.4: Big Data process life cycle

However, due to the innovation of information technology and Big Data software and techniques, organizations can overcome the difficulties. A study has presented several solutions for Big Data processing problems using Big Data technology derived from the IT eco-system based on advanced data-analysis on top of the IT Servers, System Architecture, or Network and Physical objects virtualization. The solutions cover all of the challenges related to capturing data, organizing data, analyzing data, and making value-enhancing and appropriate decisions for the concerned stakeholders (Sanyal, Bhadra, & Das, 2016). One type of Big Data software collects information based on the preferences of users and provides recommendations based on location using small-scale datasets systems. Researchers have conducted a study and analysis to examine the quality parameters of recommendation systems for location-based social networks (LBSN) with Big Data

(Narayanan & Cherukuri, 2016).

Another Big Data technology for collecting and storing social media data is NoSQL databases. A NoSQL database is a database management system (DBMS) that manages non-relational Big Data effectively and with high-performance when reading and writing. A NoSQL database is currently being used by Google, Amazon, Facebook, and many other major organizations operating in the era of Web 2.0 (Han, Haihong, & Du, 2011; Kaur, & Rani, 2013; Russom, 2013). A study has used NoSQL methods for capturing and storing data from Twitter during natural disasters. The researchers proposed and presented two different approaches for tracking social media users' activities and analyzing social media data during natural disasters focusing on Twitter users (Bruns, & Liang, 2012). Researchers addressed the challenges related to the storage of data through the Infrastructure as a Service (IaaS) cloud environment using FRIEDA (Flexible Robust Intelligent Elastic Data Management), an application specific storage and data management framework for compostable infrastructure environments (Ramakrishnan & Ghoshal, 2014).

Other kinds of Big Data technology for storing are the in-memory database and the private cloud. In-memory databases are a kind of database offering fast access and high performance. On the other hand, the private cloud is an effective data management strategy due to its allocation and reapportionment of virtualized system resources (Russom, 2013). A study has developed a model based on empirical data for assessing the benefits of using storage clouds versus purchasing disk drives (Walker, 2010).

1.1.7.2 Big Data Access

In the Big Data era, accessing a large amount of data is very challenging. The challenge is

how to access Big Data at a certain level of detail with high speed to be able to analyze the data and make decisions rapidly. In order to find meaningful information from accessing massive amounts of data in a short time, it is essential to be able to access the data at every period because some data are updated every minute. There are two suggestions to overcome this challenge, and both allow organizations to investigate a large amount of data and gain business insights in near-real time. One is to increase the memory and power of the parallel processing to access massive volumes of data very fast. The other suggestion is using a grid computing method to store data in-memory (SAS Visual Analytics, 2013).

In addition, many tools and techniques allow Big Data accessing in a reasonable period of time. The solution is a product development software services and solutions company, providing various techniques for accessing Big Data such as SQL on Hadoop, OLAP, Self-service Data Discovery, Data Virtualization, and Search (Archived Webinar, 2015). Self-service data access and visualization is an approach that allows users to access and work with data without writing SQL or MapReduce techniques; this is available only using Tableau with Impala for Big Data applications (Cloudera VISION, 2015).

1.1.7.3 Big Data Analysis

The reason for Big Data analysis is to make decisions in order to provide business insights and opportunities. There are three kinds of Big Data analysis. Retrospective data analyses involve analyzing the historical data using statistical analysis approaches to make assumptions about the future. Another kind of Big Data analysis is predictive data analyses, which generate scenarios based on historical data using simulation models to understand the future. Prescriptive data analyses are a third kind that analyzes real-time data to plan for future actions (Power, 2011). An

online article indicates that some business departments, such as retail business, consumer behavior, and performance, use Big Data analysis to understand customer movement in the stores or online on websites, as they engage in transactions, product searches, etc. (Bucholtz, 2012).

MIT Sloan Management Review, together with the IBM Institute for Business Value, has conducted a survey of about 3000 executives, managers, and analysts from more than 30 businesses and 100 countries. The purpose of the study is to understand how organizations are using analytics to increase awareness and support action. They found that the top-performing organizations rely on analytics in their activities, but some organizations are struggling in their use of the analytics approach owing to managerial and cultural reasons. Also, they stated that visualization analysis is definitely valuable for purposes of data analysis (LaValle, Lesser, Shockley, Hopkins, & Kruschwitz, 2013).

Numerous data analysis approaches and methods depend on the kind of the data. For structured data, Business Intelligence, Cluster Analysis, Data Mining, and Predictive Modeling are appropriate data analysis techniques. On the other hand, Crowd Sourcing, Textual Analysis, Sentiment Analysis, Network Analysis and Analytics 3.0 era tools are analysis methods for unstructured and semi-structured data (Gandomi & Haider, 2015). Data mining technology is used to explore patterns in large amounts of data. It includes statistical analysis, decision trees, and visual graphics to help in analyzing information for business decisions (Shaw, Subramaniam, Tan, & Welge, 2001). Text mining is one field of data mining that allows users to analyze large amounts of text data from different sources and to extract the topics and terms from the information (Hearst, 2003). Sentiment analysis is used for social media data analysis such as Twitter and is a natural language processing approach used to identify and extract subjective information (Kouloumpis, Wilson, & Moore, 2011).

For Big Data analysis, specific useful Big Data analytical tools include software such Hadoop, Python, Pig, Hive, SPSS, R, Cloudera, and SAS (Russom, 2011). Several technological approaches need to work together to get the most value from the information. For example, data management, data mining, Hadoop, in-memory analytics, predictive analytics, and text mining are techniques that work in cooperation to achieve value in Big Data analysis (SAS.com). Moreover, the use of Tableau and Impala make possible powerful visual analytics on Hadoop. Impala is an open source analytic database for Apache Hadoop, which processes data quickly (Cloudera Vision, 2015).

1.1.8 The Evolution of the Professions

With the evolution of technology and Big Data, professions have evolved and changed to address the new qualifications required for the meaningful use of Big Data in a manner that allowed for new advancements. For example, the evolving of the job function over longer periods has seen the movement from non-technical traditional employees to engineers who focus on data driven growth. As technology evolves, most businesses have increasing needs for new technical roles and requirements. These changes have resulted in changing job titles and the associated compensation in an effort to attract new and better e talent (Issid, 2017). Big Data jobs are resulting in an evolution of careers (Levine, 2011), and include a variety of different skills related to Big Data analysis (Hardin, Hoerl, & Horton, 2015). Big Data professionals typically have an advanced degree in one of several related but different fields such as statistics, applied mathematics, operations research, or economics, and business (Cleary, &Woolford, 2010; Ahern & Keller, 2014). In addition, around 86% of Big Data professionals have at least a master's degree, and 20% have a Ph.D. Nicholls (2001) suggests that future needs require adding IT-related courses that

would enhance the training of the statistical analyst profession. Also, Big Data professionals generally have experience in advanced analytics tools and methodologies, such as data mining, modeling, advanced coding, and programming.

1.2 Problem Statement

The arrival of the Big Data era has become a major topic of discussion in many sectors because of the premises of Big Data utilization and its impact on decision-making. It is an interdisciplinary issue that has captured the attention of scholars and created new research opportunities in Information Science, business, healthcare, and many others fields. The problem is the concept of Big Data is not well defined, so there exists confusion in IT regarding what jobs and skill sets are required in the Big Data area. The problem stems from the newness of the Big Data profession. Because many aspects of the area are unknown, organizations do not yet possess the IT, human, and business resources necessary to cope with and benefit from Big Data. These organizations include health care, enterprise, logistics, universities, weather forecasting, oil companies, e-business, recruiting agencies, etc., and are challenged to deal with high volume, high variety, and high velocity Big Data to facilitate better decision-making.

Because of the emergence of the notion of Big Data, many questions have been asked about Big Data, regarding what all these data mean, who gets access to what data, how data analysis is used, and how these data can be measured.

1.3 Research Question

The lack of an established definition and common terms as applied to Big Data has resulted in some different approaches, different methods, and different jobs, all of which are

involved in Big Data. This research examines the definition of Big Data, the professions that are involved and explores the influence of Big Data in one specific application to provide insight into this emerging research arena. To address this broad area of research within the context of testable hypothesis the following research questions were developed.

- How can SM data help businesses to contribute to better decision making?
- How did SM data shape the 2016 presidential election?
    - Are millennials influenced by people whom they follow on social media?
    - Did social media data use shape millennials' voting decisions?
    - Did people's attitudes toward participating in social media discussions influence millennials' political decisions?
- Within the field of Big Data domain, what are the differences and commonalities among job description elements for Statistical Analysts (SA), Big Data Analytics Professionals (BDA), Data Scientists (DS), Data Analysts (DA), and Business Analytics Professionals (BA)?
- What has been the dynamic behavior among these differences and commonalities, in recent years?

1.4 Purpose and Contribution

This research proposes a new way to look at Big Data and Big Data analysis. It is well-suited to the theoretical and methodological foundations of Big Data analysis and addresses an increasing demand for more powerful Big Data analysis from the academic researchers' perspective. Essay 1 contributes a strategic overview of the untapped potential of social media Big Data in the business world and describes its challenges and opportunities for aspiring business organizations. It also offers fresh recommendations on how companies can exploit social media data analysis to make better business decisions—decisions that embrace the relevant social qualities of its customers and their related ecosystem.

Essay 2 contributes to a better understanding of the influence of social media during the 2016 American presidential election and develops a model to examine individuals' attitudes toward participating in social media (SM) discussions that might influence their decision in choosing between the two presidential election candidates, Donald Trump and Hilary Clinton. Essay 3 contributes a clarification of the skill requirements for Big Data professionals for the joint benefit of the job market where they will be employed and of academia, where such professionals will be prepared in data science programs, to aid in the entire process of preparing and recruiting for Big Data positions. Also, it addresses the need for clarification of such overlapping subjects not only in industry but also in the academic environment.

1.5 Research Design

This study employs a mixed methodology of qualitative and quantitative research through the analysis of data collected through survey instruments, semi-structured interviews, and latent semantic analysis of online job descriptions. Essay 1 presents a strategic overview of the untapped potential of social media Big Data in the business world by describing its challenges and opportunities for aspiring business organizations. This research offers fresh recommendations on how companies can exploit social media data analysis to make better business decisions—decisions that embrace the relevant social qualities of its customers and their related ecosystem.

Essay 2 integrates social influence theory, TAM2, and media richness theory and develops a model to examine individuals' attitudes toward participating in social media (SM) discussions that might influence their decision in choosing between the two presidential election candidates, Donald Trump and Hilary Clinton. The result will provide a better understanding of the influence of social media.

The third and final essay examines the differences and commonalities among company-posted job requirements for five professions, Statistical Analysts (SA), Big Data Analytics Professionals (BDA), Data Scientists (DS), Data Analysts (DA), and Business Analytics Professionals (BA), by obtaining and analyzing online job descriptions through latent semantic analysis (LSA). The result will then be used to clarify skill requirements for Big Data professionals for the joint benefit of the job market where they will be employed and academia, where such professionals will be prepared in data science programs.

1.6 Organization of the Dissertation

This manuscript includes an overview of information, information science, and the relationship between information science and business management information systems and Information Science, beginning with a literature review of the impact of information science on the business practice of Big Data phenomena and Big Data analysis. The next section provides literature review and background on the history of Big Data. The third part presents the methodology of each essay. Results and discussions of the three essays are presented in Chapter 4. Finally, Chapter 5 is the conclusion of this dissertation research and gives suggestions regarding where the future of the Big Data research lies.

CHAPTER 2

LITERATURE REVIEW

2.1 The Definition of Big Data

The arrival of the Big Data era has become a major topic of discussion in many sectors. It is an interdisciplinary phenomenon that has captured the attention of scholars and created new research opportunities in information science, business, health care, and many other fields (Chen & Zhang, 2014). Recently, health informatics, business, management science, decision science, and many other fields have become involved with the premises of Big Data utilization and their impact on decision-making. A report shows that research into the potential of Big Data has increased among several social and behavioral scientists (Snijders & Matzat, 2012).

Data are created rapidly from a wide variety of sources by people and technology, such as satellite images, social media posts, online transactions, smart phones, online videos, academic research results, etc., showing that this is truly the era of Big Data (Zhang, Stoffel, Behrisch, & Keim, 2012). A simple example of Big Data at work is the Nike application for the iPhone or iPad. When a person works out, this application collects and tracks distance and calories burnt (Zaslavsky, Perera, & Georgakopoulos, 2013).

The term Big Data has no single definition. Mayer and Cukier (2013) define Big Data as a collection of data from traditional and digital sources inside and outside the organization, which has created a huge volume of data that can be dealt with on a large scale only. Also, Boyd and Crawford (2012) have described Big Data as a phenomenon which can be viewed from three different perspectives, those of technology, analysis, and mythology.

Other researchers (Kim, Trimi, & Chung, 2014) have defined Big Data in terms of its characteristics, volume, variety, and velocity, also known as the 3 V's and sometimes as the 4 V's.

Several researchers have considered these characteristics as attributes and challenges. The term volume relates to data size in terabytes (TB), petabytes (PB), or zettabytes (ZB), and variety refers to the types of data. Velocity refers to how often the data are produced, and the fourth V is the valuable information in the data (Zaslavsky, Perera, & Georgakopoulos, 2013).

Researchers have indicated that the Big Data concept is still unclear in the way it is used around the world and that different analytical approaches to dealing with it are needed (Wielki, 2013). A common definition of Big Data with which several researchers agree is that "Big Data is a collection of a huge data sets with a great diversity of types, so it becomes difficult to process by using state-of-the-art data processing approaches or traditional data processing platforms" (Chen, Zhang, 2014, p. 1). This is the definition that will be used for this paper. The important features of Big Data are related not only to the volume of data, but also to processing and to the acquisition of value from these data to help in decision making (Kim, Trimi, & Chung, 2014).

The size of data sets has grown rapidly. A recent study states that the total amount of data worldwide exceeded one zettabyte in 2010 and that it has exceeded 1.8 ZB since then. In addition, at current rates of growth, it is expected that by 2020 the amount of data will be 35 ZB. According to the Information and Communications Technologies (ICT) organization, this estimate may prove to be too conservative (Zaslavsky, Perera, & Georgakopoulos, 2013). In fact, about 2.5 exabytes of data have been created every day since 2012, and this amount of data is doubling about every forty months (McAfee & Brynjolfsson, 2012). Significantly, over the last few years, 90% of the historical data have been generated (Wielki, 2013) and produced from different sources.

Big Data is moving from the realm of academic research into that of everyday business transactions and encounters. For example, according to "How Big Data Analysis Helped Increase Walmart's Sales Turnover," "Walmart has changed decision making in the business world

resulting in repeated sales. Walmart observed a significant 10% to 15% increase in online sales for $1 billion in incremental revenue. Big Data analysts were able to identify the value of the changes Walmart made by analyzing the sales before and after Big Data analytics were leveraged to change the retail giant's e-commerce strategy" (2015).

2.2 History of Big Data

Big Data is not a new topic; it was introduced in the early stages of technological advancement. In the 1960s, Big Data was called data processing, which is "the collection and manipulation of items of data to produce meaningful information" (French, P1, 1996). Between the 70s and the 80s, Big Data was called information application. In the 1990s, the term information application was changed to data warehousing and mining. From the information management perspective, Big Data is similar to the phenomenon of information overload, a point that was noticed by social scientist Georg Simmel in 1950 (Edmunds & Morrise, 2000). Today, the concept denoted by the variety of original phrases from data processing to data mining is called Big Data (Kim, Trimi, & Chung, 2014).

Table 2.1: Evolution of Big Data terminology (Sharda et al., 2017)

1970s	Decision Support Systems (DSSs)	Routine Reporting AI/Expert System Decision Support Systems (DSSs)
1980s	Enterprise/executive IS	Relation DBMS On-Demand Static reporting Enterprise Resources Planning
1990s	Business Intelligence	Executive Information Systems Dashboards & Scorecards Data warehousing
2000s	Analytics	Business Intelligence Data/ Text Mining Cloud Computing, SaaS
2010s	Big Data	Social Network/ Media analytics In-memory, In-Database Big Data Analytics

From the business information system perspective, Sharda et al. (2018) book illustrates a timeline that shows the expressions used to describe analytics since the 1970s (p. 13) shown in Table 1. From Table 2.1, we can see that Big Data is the development of decision support systems (DSSs), enterprise executive information systems, business intelligence, and analytics. This study follows Sharda et al.'s illustration.

2.2.1 The 1970s Period: Decision Support Systems (DSSs)

Organizations focused on structured and routine reports that were produced from information systems to help managers in making better decisions. A routine reporting system is "a mechanism for monitoring the 'mission' of an organization" (Rainier, & Cegielski, P 234, 2012). The manager uses these reports for decision-making. Businesses create routine reports to inform managers about what has happened in previous periods (daily, weekly monthly). However, managers needed a different level of reporting to have a better understanding and facing business challenges and unstructured problem, so Decision Support Systems (DSSs) were developed.

DSSs are based on theoretical studies of organizational decision making, introduced in the early 1970s by Michael S. Scott Morton under the term "management decision systems." He defined DSSs as "interactive computer-based systems, which help decision makers utilize data and models to solve unstructured problems" (Sprague, 1980). In other words, they are computer information systems that can provide information in different analytical decision models and allow for database accessing in order to support a decision maker in making decisions effectively in complex and ill-structured (non-programmable) tasks (Klein and Methlie, 1990; Finlay et al., 1998). DSSs provide supports to business or organizational decision-making activities; services to management, operations, and planning levels in the organization; and benefits to decision makers

in problems solving stage (e.g., Unstructured and Semi-Structured decision problems) (Keen, 1987). During the late 1970s and early 1980s, there was a rapid increase in the development and application of DSSs; a new model had emerged, termed the expert system (Ford, 1985).

An expert system (ES) is a program for problem-solving that requires specialized knowledge and skill and processes the knowledge of experts, attempting to mimic their thinking, skill, and intuition (Ford, 1985). Basically, DSSs and ESs have the same goal, which is to improve the quality of the decision, but different objectives. In the case of a DSS, the objective is to help the user make decisions may making data and relevant models easily accessible. The models and the data can be assessed and used by at any time. By contrast, an ES has as its objective to always provide the user with a correct decision or conclusion. While achieving this objective is not always possible, the true standard of a satisfactory ES is performance as good as or better than that of an expert in providing decisions or conclusions to non-experts (Ford, 1985).

2.2.2 The 1980s Period: Enterprise/Executive Information Systems

The 1980s saw a significant change in the way organizations dealing with transactions data that produced from different systems (e.g., accounting, marketing, or sales, finance, manufacturing). These systems were integrated and called enterprise resources planning (ERP) systems and were organized on rational database management (RDBM) systems (Sharda et al., 2017). Enterprise resource planning (ERP) is software with integrated applications that manage business process related to technology, services and human resources (Kimball, & Ross, 2011). ERP systems often are installed on relational databases, which consist of one or more relations in two-dimensional (row and column) format. Rows are called tuples and correspond to records; columns are called domains and correspond to fields (Kimball, & Ross, 2011). Organizations

became in need of RDBM and ERP systems when data integrity and consistency became an issue (Sharda et al., 2017). ERP involves the collection of all available data and its integration into a single system so that it can be accessed at any time by any part of the organization. In addition to the emergence of ERP systems, business reporting became an on-demand as-needed business practice. On-Demand Static reporting can be run whenever needed by the user, and the data are stored in the Completed Reports module. There are two types, Static SQL reports and Static multidimensional expression (MDX) reports. While the former is run asynchronously so that Commerce Server Business Desk can be used while the reports are running, the latter is run synchronously, so Business Desk cannot be used while the reports are running (Sharda et al., 2017).

2.2.3 The 1990s Period: Business Intelligence

Executive information systems (EISs) were developed in the 1990s in response to the need for reporting that would be more versatile (Sharda et al., 2017). The Executive Information System (EIS) is also referred to with the term Executive Support System (ESS). It is a management information system for the purpose of providing information for needs of senior executive decision making by allowing for easy access to information both inside and outside the organization that is important for the goals of the organization (Power, 2002).

DSSs are also for the purpose of facilitating the decision making of executives. The systems use graphical scorecards and dashboards that are visually appealing, and they focus on crucial factors, helping those who make decisions have access to key performance indicators. Data warehousing (DW) was soon adopted by medium to large-sized businesses as a solution for making decisions on the organizational scale. The data for the scorecards and dashboards come

from DW, and consequently, the efficiency of the ERP business transaction system was not reduced (Sharda et al., 2017)

Scorecards measure performance and compare it to goals and projections. On the basis of key performance indicators (KPIs), the success of the performance is evaluated. For management to evaluate the progress effectively, it is necessary to determine the KPIs at an early stage (Nagle, 2016). Dashboards consist of numerous reports, so the user can compare or contrast these reports and have easy access to numerous sets of data in one location. Scorecards as well as other kinds of reports can be seen on the dashboard. Exception reports, 52-week profit analyses, and new item trends are among the types of reports most frequently viewed on a dashboard. Users can customize dashboards so that different views are presented. It is best if all the data are taken from a single repository as this results in more accurate reports (Nagle, 2016). The possibility of such versatility in reporting along with the need to maintain the transactional integrity of the business information system required the development of data warehousing (DW), a middle data tier and repository that is designed to support reporting and decision making by businesses.

Data warehousing was originally termed the Decision support system (DSS) (Kimball, & Ross, 2011); it is also known as the enterprise data warehouse (EDW). EDW is comprised of an organization's areas of data warehouse staging and presentation. The EDW has been described as an atomic, centralized, and normalized data warehouse layer of the data warehouse, but it has not been clear whether this kind of system can be employed for drill-down and querying by the end-user. We prefer to think of the EDW as the greatest overall combination of presentation and staging services rather than adopting the prior interpretation of an atomic, centralized layer (Inmon, 2005; Kimball & Ross, 2011).

2.2.4 The 2000s Period: Analytics

In the 2000s, the DSSs, driven by Data Warehousing, were referred to as BI systems (Sharda et al., 2017). Business intelligence (BI) refers to the leveraging of the organization's external and internal information assets to facilitate better decision making (Kimball, & Ross, 2011). Along with the increasing accumulation of longitudinal data in the DWs came increased software and hardware capabilities to accommodate decision makers' evolving needs (Sharda et al., 2017).

The globalized and competitive marketplace requires decision-makers to have information in a format that would make it digestible so that they can solve business problems and take advantage of opportunities in the market (Sharda et al., 2017). Data Warehouse data do not reveal the most recent information because they are updated only periodically. This latency problem can be reduced through a subsequently developed system of more frequent data updating. The result is referred to as real-time data warehousing, or, more realistically, right time data warehousing. The difference between these two is that the latter uses a policy for data refreshing that depends on how fresh certain items need to be (not all of the items of data have to be refreshed on a real-time basis) (Sharda et al., 2017). Data Warehouses have a large volume and many features. It became necessary to use them to "mine" the data of companies in order to "discover" useful specific items of knowledge to ameliorate business practices and processes; this led to the use of the terms text mining and data mining (Sharda et al., 2017).

Data mining consists of undirected queries. It often queries data that are very atomic, and it looks for data patterns that are unexpected (Fayyad et al., 1996). Clustering, estimating, classifying, predicting, and co-occurrence patterns are all considered valuable results. Data mining uses a variety of tools, including neural networks, decision trees, memory- and case-based

reasoning tools, genetic algorithms, visualization tools, and fuzzy logic, in addition to classical statistics. Data mining is usually a data warehouse client (Fayyad et al., 1996; Kimball, & Ross, 2011).

The purpose of text mining is to cope with the problem of information overload by means of techniques arising from machine learning, data mining, information retrieval, natural language processing, and knowledge management. Texts are mined through document collection preprocessing (text categorization followed by information and term extraction), storing and analysis (including clustering, distribution analysis, association rules, and trend analysis) of the intermediate representations, and visual presentation of the results (Feldman, & Sanger, 2007).

Greater volumes and more types of data required more processing power and more storage. While larger corporations had the necessary resources to cope with these issues, a more financially practical business model is required for smaller businesses. This need led to the development of the business model. As a result, these smaller businesses found ways to access these analysis services as needed rather than purchasing the hardware and software resources that would be generally productive (Sharda et al., 2017).

One way to have easy network access to share and configurable computing resources such as servers, storage, networks, services, and applications without the need for much interaction with the service provider is through cloud computing (Mell, & Grance, 2011). Through Software as a Service (SaaS), the consumer can use the applications made available by the provider in the cloud. These applications can be accessed through various kinds of devices by means of a thin client interface, like a browser, or by means of a program interface. The only control or management required of the user of items such as servers, network, storage, operating systems, or application capabilities is certain configuration settings for applications (Mell, & Grance, 2011)

Organizations using SaaS do not need to install applications or run them on their computers or data centers (Rouse, 2016). The 2010s are witnessing another change in the capture and use of data. High levels of internet usage have resulted in new mediums for data generation. There are several new data sources, including digital energy meters, tags, radio-frequency identification (RFID), clickstream, smart home devices, weblogs, equipment worn by users to monitor their health, and smart home devices. The most interesting of the new data sources is social media for social networking.

While such unstructured data offer rich information content, its analysis is challenging for the computation system in terms of both hardware and software issues. The use of the term Big Data highlights the challenges of analyzing these new data sources. Developments in hardware (including large memory resources, massively parallel processing, and parallel multiprocessor systems) and software and algorithms (for example, NoSQL and Hadoop with Map Reduce) have been aimed at meeting these challenges.

Social media analytics exists for the purpose of creating and assessing informatics frameworks and tools to collect, analyze, monitor, visualize, and summarize social media data, to make interactions and conversations easier, and to find patterns and intelligence that can be used effectively (Zeng et al., 2010). A great amount of social network data are produced by social media, and it is possible to mine these data in search of applications for business purposes. With data mining techniques, researchers and others who use such data can analyze social media data that are large scale and complex and that often change. The text contained in the nodes of social network takes a variety of forms. Examples include news articles, blogs, and links to other posts. It is possible for users to tag each other, and such tags should also be considered a type of text

data. The quality of the inferences the user makes about graphs as well as social networks can be enhanced through using content (Aggarwal, 2011).

Another kind of database management system is referred to as the In-memory, In-Database system. It is an in-memory database (IMDB) and can also be described as a main memory database (MMDB) system or memory resident database, and it makes primary use of the main memory of the computer, rather than a disk, for storage of data. The In-Memory Data Grid functions in conjunction with a database that already exists. It provides a layer of in-memory storage and processing between the application and the database that is massively distributed. This layer provides super-fast access to and processing of data for applications.

2.3 Big Data Sources: Social Media Sites

Social media is defined as "forms of electronic communication (as Web sites for social networking and microblogging) through which users create online communities to share information, ideas, personal messages, and other content (as videos)" (Schauer, 2015; Shah, 2017). In addition, social media refers to websites (e.g. Facebook, Twitter, Instagram, etc.) that allow individuals to create share, or exchange information and ideas with other people and businesses about different topics and aspects in different forms, image, video or text (Fleck et al., 2015; Shah, 2017). Social media has replaced the traditional media and the industrial media because of its quality, speed, performance, and frequency (Differencebtw, 2015).

Social networking platforms are a third-party place where individuals and organizations communicate with others and exchange information and ideas about many aspects (Erlandson, 2013). Social networking platforms can be on any social media. The main purpose of social networking is engagement with others, creating relationships and building up of trust and loyalty

because it is reliable and support a wide range of interests and activities. It is important for the users to exchange information (Burke, 2013). Social network platforms include Facebook, Twitter, Tumblr, Instagram, LinkedIn, YouTube, Yelp, etc. (Shah, 2017).

The main differences between social media and social networking sites are as follows: 1) in social media, the information is shared widely, and all users can access and share information equally, while social networking is for interaction between people and organizations who have common goals or interests; 2) social media is entirely virtual and depends solely on the Internet, while social networking can take place via social media or can occur in an actual physical community; 3) while a person can communicate a message to the public via social media, messages sent through social networking go only to those who are part of the group; 4) social media is useful for business, while social networking is more useful for having discussions and creating relationships; and 5) the news on social media is often based on rumor and lacks a dependable source, while the news sent through social networking tends to be reliable and authentic (Aggarwal, 2011; "Differencebtw," 2015; Schauer, 2015).

2.4 The Reasons for Studying Big Data from Three Different Approaches

Big Data is an emerging area not only because data is being collected automatically at a faster rate than ever before but also because we are on the verge of developing tools that make the data more accessible and more usable. This dissertation approaches this area of research from four different approaches. The first establishes the rise of Big Data via a historical approach which is discussed in the literature review. The second approach examines how business can benefit from learning to use and leverage new sources of data such as social media data. The third approach includes a survey of social media users to study the influence social media has in their decision

processes. The last approach examines the rise of the Big Data profession by examining the jobs that industry is attempting to fill.

The purpose of the first study, the conceptual research, is to describe how SM data can contribute to better business decisions—decisions that embrace the relevant social qualities of a business' customers and their related ecosystem. More specifically, this research 1) provides a strategic overview of the untapped potential of SM data in the business world, 2) describes SM data challenges and opportunities for aspiring business organizations, and 3) provides fresh recommendations on how companies can exploit social media data analysis to make better business decisions.

The goal of the second research study is to investigate social media data's influence, which can help to explain the biases reflected in the millennials' decisions. This study seeks to understand better the impact of social media influence in decision making and because of its general applicability to many groups uses the 2016 U.S. presidential election as the venue for testing such influence. Examination of the potential attitudes surrounding the use of SM discussion platforms and how they influence individual decisions is difficult because of the varied interests of different market segments.

The goal of the third research study is to clarify skill requirements for Big Data professionals for the joint benefit of the job market where they will be employed, and academia, where such professionals will be prepared in data science programs. We pursue our goal by examining differences and commonalities among company-posted job requirements for the five professions listed in the previous paragraph. In accomplishing this goal, we hope to contribute to the improvement of the entire process of preparing and recruiting people for Big Data positions,

and address the need for clarification of such overlapping subjects not only in industry but also in the academic environment (Mauro et al., 2016).

The methodology that was used to answer each of the three research questions is discussed in the next section.

CHAPTER 3

METHODOLOGY

3.1 Essay 1 Value of theoretical foundation based on literature review

A traditional literature review was conducted in October 2016. We systematically searched the following specialized database sources: Business Source Complete, EBSCOhost, and Business Insights: Global, and Business Research Databases. Additionally, Google Scholar searches were performed. Search words and phrases included Facebook Data, Social Media Big Data, Social Media Data, Social Media, Twitter Data, and Business in Social Media Applications. The ‘‘snowball’’ method of using the most recent works to find relevant articles cited in them provided additional articles. Because keywords in research articles are not based on common lists, it is certainly possible but unlikely that a significant literature contribution was missed.

3.2 Essay 2 Survey Research

After the 2016 election, the researchers for the current study conducted an online survey of college students at a large public university in the southwestern U.S. The focus of this study was on millennials' use of social media (SM); specifically, undergraduate college students who were enrolled in business classes and political classes were considered an appropriate population. Research indicates that college students rely much more than older adults on SM as a source of political campaign news (Pew Research Center, 2016).

The authors developed a survey instrument by adapting established measures from prior studies. Measures of social media norms and social media community identification were all modeled and adopted from previous research and contextualized for this research (Hsu & Lin, 2008). TAM2 constructs (usefulness and ease of use) were operationalized and measured using

items adapted from other research and used author-developed scales (Venkatesh & Davis, 2000). Measures for perceived social influence were adopted from Carlson et al., (1999). The measure of Attitude was adopted from Webster and Trevino's (1995) study. For a complete list of measures, see Appendix A. Seven-point Likert scales were used to capture student responses.

This research proposed hypotheses about how participation in SM discussions affects individuals socially and influences their attitude in making voting decisions, in an effort to explain the unexpected outcome of the 2016 U.S. presidential election in the context of social influence theory showing that participating in social media sites influences individuals' attitudes in making voting decisions (Varnali et al., 2015). Based on previous work on social influence, media richness, and SM influence, the authors of this study defined the key constructs, developed hypotheses, and put forward a conceptual framework, as shown in Figure 3.1.

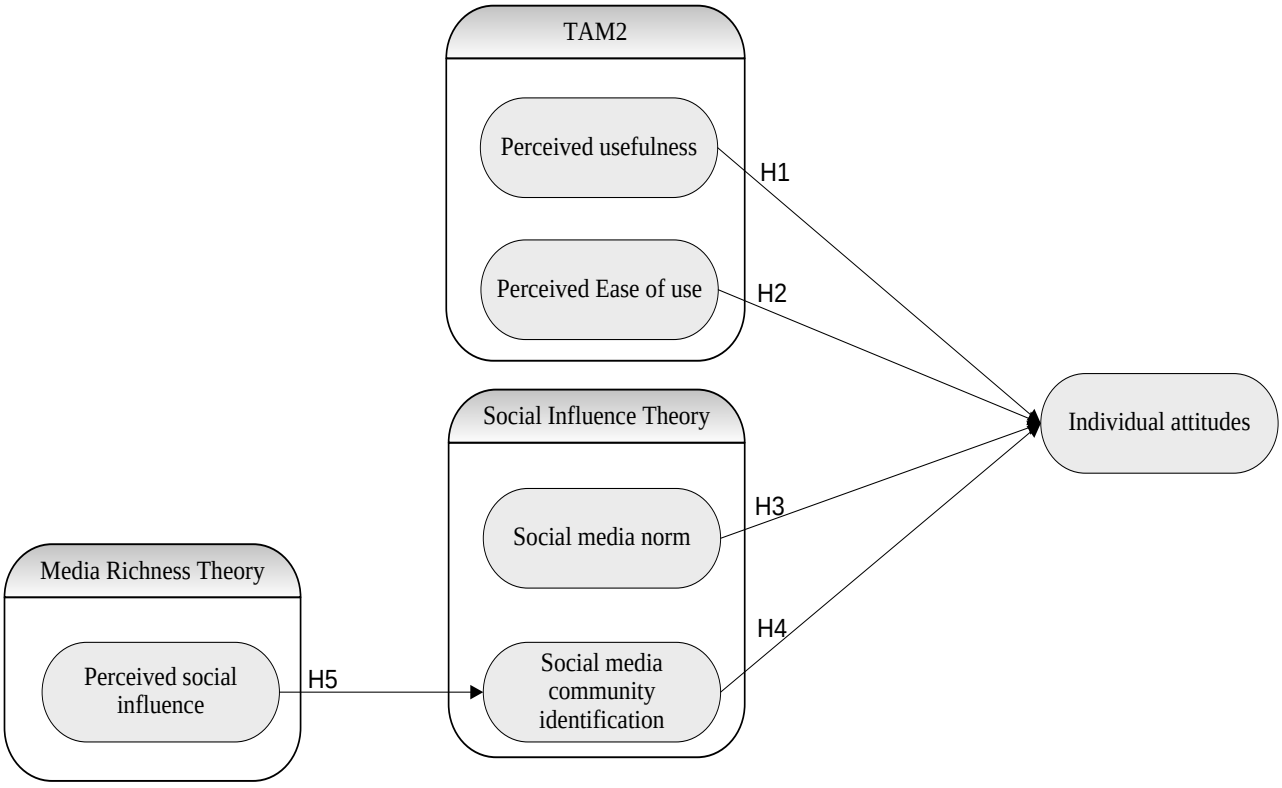

Figure 3.1: Conceptual framework

3.2.1 Perceived Usefulness and Perceived Ease of Use

These two constructs are drawn from TAM2. This study adopted perceived usefulness and perceived ease of use about social media from Porter et al., (2006) and from Hsu & Lin (2008). Perceived usefulness is the degree to which an individual believes that using an SM platform influences his or her performance in making decisions. Perceived ease of use is the degree to which an individual believes that using an SM platform is free of effort. The more that an individual perceives the SM platform as useful and easy to use, the more favorable that individual's attitude toward the use of SM discussion platforms (Porter et al., 2006). Thus, we propose the following hypothesis:

H1: There will be a significant positive relationship between perceived usefulness and individual attitudes.

H2: There will be a significant positive relationship between perceived ease of use and individual attitudes.

3.2.2 Social Media Norm (SN) and Social Media Community Identification (CI)

The social influences theory provides a theoretical basis for a relationship between social norms and individuals' attitudes (Venkatesh et al., 2003). This study adopts the social norm concept from Hsu & Lin (2008). It defines social media norms as related to the level at which an individual recognizes that his/her choices and attitudes are endorsed by others, through participation in SM discussions. Empirical research indicates that individual attitudes about participation in SM discussions influence their voting decisions (Silver et al., 1986). An individual's attitude toward participating in SM discussions with others that share the same norm can influence the individual's voting decisions. Also, this study adopts the community identification in social media concept from Hsu and Lin (2008). Community identification in social media sites leads to a sense of belonging to a particular group among members of an SM discussion

platform. An individual's attitude toward participating in SM discussions with the group to which the individual belongs can influence the individual's voting decisions. Thus, we propose the following hypotheses:

H3: There is a significant positive relationship between social media norms and individual attitudes.

H4: There is a significant positive relationship between community identification and individual attitudes.

3.2.3 Individual Attitudes

This study adopts the individual attitudes concept from Hsu et al., (2008). Individual attitude is the preference for participating in SM discussions, which may have influenced the individuals' decisions to vote for a 2016 U.S. presidential candidate. Use of SM sites in general influenced people's attitudes socially and politically (Moy et al., 2005; Shah et al., 2001; Wellman et al., 2001).

3.2.4 Perceived Social Influence

Perceived social influence is a construct drawn from media richness theory; it is the change in an individual's thoughts, feelings, attitudes, or behaviors that result from interaction with another individual or a group (Rashotte, 2007). For this study, we adopted the perceived social influence concept from Carlson et al., (1999). An individual can be influenced socially by his/her group's posts such as discussions, image, videos, etc., on SM sites. We propose the following hypothesis:

H5: There is a significant positive relationship between perceived social influence and social media community identification.

After receiving approval from the university's institutional review board, the researchers approached instructors, who posted a link to the survey on course websites, and administered the survey online. All students were offered extra course credit to encourage participation. Students in a total of 10 classes were asked to complete the survey. The authors received a total of 1,101 responses, including 195 from international non-voters. After cleaning the data to eliminate the unusable responses, including those that indicated a lack of variance (i.e., from respondents selecting all 1's or all 7's) and incomplete surveys, 450 usable responses remained for further analysis, resulting in a 40% response rate. The sample achieved one of the goals of the research, that of targeting younger participants; 68% of the respondents were under the age of 21. Most participating students were male (52%). Most students (55%) had voted for Hillary Clinton, and 27% of the students had voted for Donald Trump. Complete survey demographics are provided in Table 3.1.

Table 3.1: Respondent demographics

Gender			Age			Voting		
Male	232	51.56%	18-21	308	68.44%	Donald Trump	123.00	27.33%
Female	216	48.00%	22-25	85	18.89%	Hillary Clinton	248.00	55.11%
Others	2	0.44%	26-29	27	6.00%	Gary Johnson	32.00	7.11%
Academic Statues			30-33	13	2.89%	Jill Stein	11.00	2.44%
Freshman	118	26.22%	34+	17	3.78%	Other	36.00	8.00%
Sophomore	106	23.56%	Hillary Clinton voters			Donald Trump voters		
Junior	129	28.67%						
Senior	72	16.00%	Female	141	31.33%	Female	42	9.33%
Graduate	25	5.56%	Male	106	23.56%	Male	80	17.78%

3.3 Essay 3 Latent Semantic Analysis

3.3.1 Data Collection

Data were collected from the online source http://jobs.monster.com at three points in time: August 2015, July 2016, and July 2017. We searched for job listings, focusing on the five Big Data professions (SA, DS, BDA, DA, and BA) described in the previous sections. The geographical location was set to retrieve job openings anywhere in the USA. The hiring companies/organizations belonged to a variety of industries. Each job description was pulled from the website, and a document library was created. Three hundred job descriptions were collected each year, for a total of 900 job descriptions.

3.3.2 Text Analytics

The collected job qualification records were analyzed using Latent semantic analysis (LSA), which is a text analytic method that identifies text usage patterns to simulate word meaning (Deerwester et al. 1990; Landauer & Dumais, 1997). The analysis followed the guidelines in Evangelopoulos et al. (2012) and the steps in Kulkarni et al. (2014), which are listed below.

In step 1, we compiled a term-by-document frequency matrix, also known as the vector space model (Salton 1975). This is a data structure that quantifies unstructured text, by recording the frequency of each term in each document. In order to finalize the set of documents and account for job description content effectively, job descriptions were split into individual passages that correspond to paragraphs, sections, or list items, likely to cover about one topic each. The resulting data set consisted of 8,986 such individual passages of unstructured text, which represent the columns in the term frequency matrix. The average passage size was 86 characters. In order to finalize the set of terms, we excluded the terms that appeared only once in the entire collection, as

well as trivial English words (stopwords), such as and, the, therefore, etc. The spelling of terms that appear in different spelling styles was standardized, and acronyms were spelled out (e.g., BS to Bachelor's degree). The final list of terms included terms that appear in at least four passages in the entire data set. These terms were conflated with term stemming, to merge terms that share a common stem, for a total of 2,519 unique stemmed terms. The raw term frequencies were transformed by applying an inverse document frequency (TF-IDF) term frequency transformation function. This weighting method discounts the occurrence of high-frequency terms and promotes the occurrence of low-frequency terms, making it easier to identify abstract concepts (Evangelopoulos et al. 2012).

In step 2, the transformed term frequency matrix A was subjected to singular value decomposition, or $A = U\Sigma V^T$, where U is the term eigenvectors, V is the passage eigenvectors, and Σ is a diagonal matrix of singular values (i.e., square roots of common eigenvalues between terms and passages).

In step 3, a scree plot was used to examine the eigenvalues, in order to decide how many topics to extract. The scree plot is shown in Figure 3.2. One obvious choice would be to extract three very broad topics. However, since the intention was to obtain some additional information, we opted for the next "elbow point" of the scree plot, at nine topics. The log-likelihood ratio test for dimensionality detection (Zhu and Ghodsi, 2006), performed on eigenvalues 4 to 23, verified $k = 9$ as the dimensionality estimate (observed value of the test statistic $Q_n = 37.68$, p-value = 0.0022).

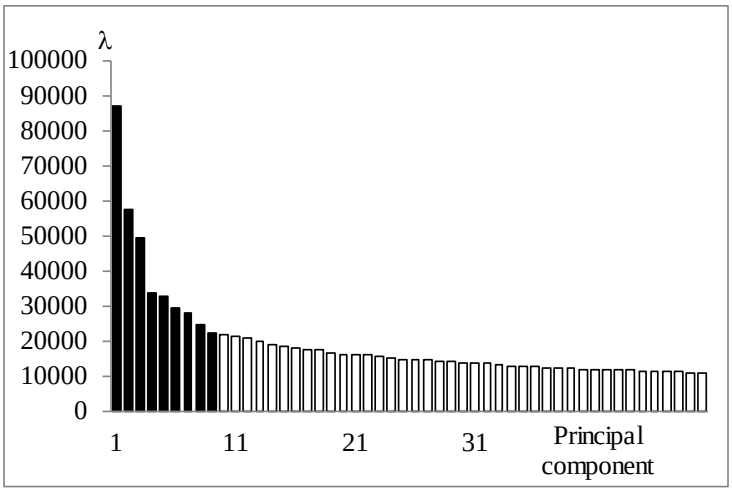

Figure 3.2: Scree plot of eigenvalues

In step 4, we labeled the topics using high-loading terms and high-loading documents. Topic labels and related high-loading terms and documents are shown in Table 3.1. Two of the authors compared their topic labels and reached consensus quickly and without controversy.

CHAPTER 4

RESULTS AND DISCUSSION

4.1 Essay 1: Using Social Media Data Contributes to Better Business Decisions

Insights gained from analyzing the data generated by social media can bring future customers closer to the targeting businesses and thus contribute significantly to better business decisions. Social media is an exceedingly popular, but not exclusively social, medium that produces vast quantities of data potentially invaluable to businesses who wish to understand their customers. The so-called Big Data of social media is rich with information about user demographics, opinions, image, preferences, influence, behavior, but also about business, politics, entertainment and other characteristics that permeate the Internet.

Business decision makers seek to transform the hidden insights gleaned from assimilating the essential, distinctive characteristics that are imbued within social media data into loyal customers, superior profits, and business success. This research contributes a strategic overview of the untapped potential of social media Big Data in the business world by describing its challenges and opportunities for aspiring business organizations. This research offers fresh recommendations on how companies can exploit social media data analysis to make better business decisions—decisions that embrace the relevant social qualities of its customers and their related ecosystem.

4.1.1 Introduction

When data from billions social media (SM) users are integrated into the global business context, insights gained from their analysis can contribute significantly to better business decisions. This research describes how SM Big Data (BD) can make that happen. As a relatively

new and typically-underutilized data source, SM BD now can be processed to provide fresh customer perspectives from personal, socially-oriented data. Businesses can use this new information to revise and tune existing strategies, helping them achieve their goals sooner and more efficiently (Hu et al., 2013).

Research supports these assertions. Approximately 82% of companies indicated that SM BD had become a significant resource for all levels of corporate planning and management (Starr, 2013). It contains wide-ranging, rich information sources, such as live customer restaurant-experience data, sports-event fan feedback data, current vehicle location data, on-site service response data, instantaneous inventory, order, and cost data, on-the-air political perceptions data, and many others (Manyika et al., 2011; Davenport & Dyché, 2013; Chen & Zhang, 2014). According to the 2014 Accenture Big Data Study, 93% of companies having revenues in excess of $250M rate BD initiatives as either “extremely important” or “important” (Olavsrud, 2014). Businesses tap the contemporary SM resource to stay competitive and even gain an advantage.

SM users generate huge quantities of data; the more than two billion active users post information that is current, personally descriptive, and abundant (Constine, 2017). Although institutions aggregate information in structured databases at tightly-controlled organization server sites, SM users typically distribute un-aggregated, unstructured, and generally unfocused data freely throughout the Internet ecosphere. When collected and analyzed appropriately, SM BD comprises many primary, rich data sources that contain personal data, political preferences, business wisdom, opinion leadership, influencer information, word-of-mouth effects, and even consumer behavior data—all of which may be used to improve business performance (He et al., 2015). As BD analytics becomes ever more reliable, SM user data becomes increasingly valuable for informing business decisions.

Much research has dealt with formal, structured data in the business ecosystem (Shigemi, 2001). Still, there is a need to describe the utilization of informal, unstructured SM BD and its application in the business world because some businesses still do not trust social media as a return on investment (ROI) source (York, 2016). Aggregating and analyzing unstructured data presents many challenges and opportunities for businesses. In this research, the term social media Big Data is used synonymously with social media data (i.e., SM data). To that end, the purpose of this research is to describe how SM BD can contribute to better business decisions–decisions that embrace the relevant social qualities of a business' customers and their related ecosystem. More specifically, the contributions of this research are threefold. This research 1) provides a strategic overview of the untapped potential of SM BD in the business world, 2) describes SM BD challenges and opportunities for aspiring business organizations, and 3) provides fresh recommendations on how companies can exploit social media data analysis to make better business decisions. The format of the remainder of this paper is as follows. We first provide a definition and classification of social media data. We then review the literature of social media data analytics, and the opportunities and the challenges of social media data. Then based on this exploration, we develop a set of four recommendations for how business can exploit the untapped potential of social media data—using social networking sites, targeting audience, optimizing the accounts, and finding ways in which companies can use social media data efficiently. We discuss conclusions in the last section. Finally, we describe limitations and proffer suggestions for future research.

4.1.2 Literature Review

4.1.2.1 The Definition of Social Media Data (SM Data)

This research uses the terms social media data (i.e., SM data) and social media Big Data

synonymously. SM data is defined as the collected information from social networks that show how users share, view, or engage with expressions and content of other users and their profiles (York, 2016). SM data is an assortment of large-scale, unstructured data stores produced by human activity. Examples include opinions, shares, likes, mentions, impressions, hashtag usage, URLs, hyperlinks, keywords analysis, new followers, and comments (Leskovec, 2011; Schrecl & Keirn, 2013). This definition communicates the view that SM data is a source of human behavioral data that can reveal how people communicate with, interact with, and influence each other online.

4.1.2.2 The Dramatic Increase in SM Data

Social media is the leading source of the vast increase in Big Data (Gruzd, 2016). Estimates place the number of SM users worldwide from between 2010 and 2021. By 2019, there will be around 2.77 billion SM users on earth, up from 2.46 billion in 2017 (Constine, 2017). Research indicates that Twitter users produce 200 million tweets daily (Tang, 2012), and 3.2 billion images are shared each day (Smith, 2016). Massive amounts of unstructured data are produced by Facebook users. For example, in 2009, over 600 million active Facebook users spent more the 9.3 billion hours every month on the site, creating approximately 90 pieces of content each. This content included photos, notes, or links. In 2011, the total number of Facebook users increased to one billion (Manyika, Chui, &, Brown, 2011). Every minute, users upload 24 hours of video content to YouTube (Wielki, 2013). SM sites have increased the different types and combinations of data.

SM classifications are as follows: 1) social networks (e.g., Facebook and LinkedIn), 2) blogs (e.g., Blogger and WordPress), 3) microblogs (e.g., Twitter and Tumblr), 4) social news (e.g., Digg and Reddit), 5) social bookmarking (e.g., Delicious and StumbleUpon), 6) media

sharing (e.g., Instagram and YouTube), 7) wikis (e.g., Wikipedia and Wikihow), 8) question-and-answer sites (e.g., Yahoo! Answers and Ask.com), and 9) review sites (e.g., Yelp, TripAdvisor) (Barbier & Liu, 2011; Gundecha & Liu, 2012; Gandomi, 2015).

Today, there are more than 60 worldwide SM sites (e.g., Facebook, Twitter, Pinterest, Instagram and LinkedIn, Google Plus, Snapchat, YouTube, Yelp, etc.) (Jamie, 2017). Of these sites, Facebook is the major SM platform in terms of user population and usage (Pew Research Center, 2017). SM platforms produce vast quantities of user- and system-generated data. What follows in the next section is a description of a strategy that can help the SM-data-using targeting businesses understand the nature of their data and give an idea of how they can benefit from it.

4.1.2.3 The Analogy of the SM Donut Strategy

The so-called donut strategy is a helpful analogy for understanding the many flavors of SM data structures. The donut strategy is an SM business marketing approach that describes how a single SM concept may be shared when the users perform various social messaging activities. It helps any business understand each SM data site and how the user can post the same topic differently based on the SM platforms environment, as shown in Table 4.1 (Shannon, 2015). Table 4.1 lists some examples of SM data, based on the donut strategy business marketing approach, posted in Add This Blog. Based on these examples, businesses can acquire a better understanding of each SM data site and build their own SM interpretive strategy to improve decision making.

As described in Table 4.1, Facebook data are personal SM information (e.g., feeling, experience) about a product or service. Twitter is microblogging that allows users to share current information about consuming products or service. Youtube allows users to create and upload their videos and watch other videos (e.g., product or service reviews, overviews, tutorials, interviews)

and comment on, rate, and share them. Pinterest data are descriptions or images of content or products that link to the original website. Yelp data presents a range of positive to negative reviews of service-oriented businesses at a particular location. Instagram data displays images with comments about an associated product or service. Snapchat data shows short videos and photos about event information with a personal perspective and users daily life activities. LinkedIn data provides a repository of professional employment-related information (e.g., skills, experience, qualifications). SM data analytics helps make sense of this wealth information.

Table 4.1: Analogy of the donut strategy for several SM platforms

Social Media Sites	User's Post	Explanation
Facebook	I like donuts	Users share products and content that resonates with them.
Twitter	I am eating a #donuts	Users tweet to talk about and interact with brands about their products or services.
Youtube	Here's a video of me eating donut	Users watch, create, and upload their own content and comment on, rate, and share content.
Pinterest	Here's a good donuts recipe	Users pin products and content they like from the Internet so they can come back to it later.
Instagram	Here's a photo of my donut	Users can share images they take automatically, as well as like and comment on images from people and brands they follow.
Yelp	You will like the donuts at this place	Users leave reviews of businesses they have interacted with, both positive and negative.
Snapchat	Everyone but me is at this donut festival #fomo	A quick way for users to send "snaps" of their daily life, or their perspective at an event.
LinkedIn	My skills include donut eating.	Users share their professional skills and connect with business leaders.

4.1.2.4 The Importance of SM Data Analytics

Big Data analytics is a powerful tool that decision makers use to utilize hidden insights gleaned from assimilating the essential, distinctive characteristics from large amounts of data. Although these insights may be imbued deep within the fabric of social media, they can be

extracted through purposeful analytics, thus providing a powerful mechanism for changing the way companies function (Wlodarczak et al., 2015). To further illuminate these concepts, we will characterize Big Data analytics.

Big Data analytics is defined as the process of examining large and varied data sets to explore data hidden patterns, unknown correlations, market trends, customer preferences, and other useful information that can help organizations make more-informed business decisions. Big Data analytics is similar to traditional data analysis in using statistical methods. However, Big Data analytics relies on special technique for different kinds of data (Chen et al., 2014). In fact, there is no single technology or technique that adequately embodies Big Data analytics, which is a combination of multiple technologies and techniques for processing and analyzing massive amounts of data. Examples include data management, SM analytics, data mining, in-memory analytics, and text mining (Troester, 2012; Gandomi, 2015).

SM analytics refers to the analysis of structured and unstructured data from social media channels (Gandomi, 2015). A Recent study indicates that SM data analysis (e.g., text analytics) helps predict major future events in terms of individuals' and businesses' decisions for different purposes such as the anticipated outcomes of elections, the development of financial indicators, projected box office revenues, and potential disease outbreaks (Gruhl et al. 2005; Bollen et Al., 2011; Schoen et al., 2013; Kallus, 2014; Wlodarczak et al., 2015).

For example, one empirical study uses Twitter network analysis to investigate the relationship between responses of Republican and Democratic supporters in the 2012 U.S. presidential election and best-predicted result based on the media. The media did not discuss them in the same way; the authors observed that the news media referred to different election issues in discussing Obama and Romney. Interestingly, these media discussions correlated closely with

how the candidates' supporters spoke of their preferred political candidates. The result shows the media and their different treatment of the candidates significantly influenced the election outcome (Vargo et al., 2014). In the 2012 election, Big Data provided a predictive advantage, and it is clear that more opportunities will be created for businesses shortly (Manyika et al. 2011; Chen & Zhang, 2014). However, as with most tools—especially new and powerful ones—misuse and misinterpretation are possible.

The 2016 U.S. presidential election between Republican Party candidate Donald Trump and Democratic Party candidate Hillary Clinton is an eloquent example. In this case, differential media treatment of the two candidates did not have the same effect as in the 2012 election. The relative importance of data sources and their interpretive formula had changed: Trump chose to bypass the traditional mainstream media and use SM, mainly Twitter, and Facebook, as his dais (Markman, 2016). Failure to correctly account the effects of Trump's "Twitter factor" (MSNBC, 2017) and demonstrably deficient sampling methods (Jasonkarpf, 2016) contributed to a flawed analysis by the media (Barbaro, 2016) that ultimately resulted in extensive, incorrect predictions of the election outcomes (Siegel, 2013; Piatetsky, 2016). Thus, the promise of SM data comes with a caveat because it is another, multi-dimensional, and comparatively new ingredient in the SM data analytics cake. However, with proper attention to these issues, opportunities abound for utilizing SM data, as discussed in the following section.

4.1.2.5 The Opportunities of SM Data

SM data provides businesses with opportunities to glean new understanding from the vast, largely unstructured SM repositories. Opportunities include improving customer service (Manyika et al. 2012; Chen et al., 2014). For example, Nestlé used SM data to proactively engage customers

in the market, using its central IT command center. The company had expressed dissatisfaction with its inadequate grasp of customer attitudes and desires, which were obtained through surveys and customer sampling. Nestlé sought to improve customer experience by engaging them through new information extracted from SM data. The company established a team to operate its 24/7 analytics center, charged to monitor and respond to SM conversation about its products. As a result, Nestlé improved in the Reputation Institute's index of the most reputable world companies, from 16th to 12th (Hinchcliffe, 2012).

T-Mobile used SM data to reduce customer defections by 50 percent in a single quarter. During the early 2010s, the company was withering under stress caused by customers defecting to other carriers—a challenge exacerbated by the company's lack of accurate and scalable information. Embracing SM data, T-Mobile integrated 33 million customer records, billing data, web logs, and SM data, using real-time analytics across its IT systems infrastructure to stem the outflow (Hinchcliffe, 2012). Another company, Wells Fargo, employs SM and customer relationship management (CRM) functionaries to communicate with its customers, to connect customers with each other, to guide them toward its financial products, and to acquaint them with risk management practices (Salesforce,2017).

Companies recognize that potential, but substantial, benefits to all phases of business planning can be realized from embracing SM data (Prewitt, 2017). From a management opportunities perspective, SM data can be leveraged to achieve strategic and tactical business advantage.

4.1.2.5.1 Strategic and Tactical Use of SM Data

Strategic use of SM data focuses on long-term benefits (exceeding one year), such as

increasing sales, identifying and developing new products and services, and securing new customers and markets, to retain and grow a customer base (Manyika et al. 2012; Chen et al., 2014). Dunkin' Donuts aspires not just to sell products but to establish long-term customer loyalty. The company envisions becoming an integral part of their customers' daily lives, asserting that as a result of their taking care of customers, their customers will take care of the Dunkin' brand. Dunkin' employs specialized SM software to monitor and participate with customers in conversations about their brand, thus enabling the company to drive word-of-mouth discussions, build brand recognition, and generate customer loyalty. Thirteen million Facebook fans and increased sales have led them to proclaim success (Prewitt, 2017; Salesforce, 2017; Costello, 2017). While increasing sales realize strategic benefits, directing potential customers to the sales SM page achieves tactical returns (Prewitt, 2017).

Tactical use of SM data focuses on more immediate benefits (week-to-week, day-to-day) acquired as a result of utilizing strategic strengths (Prewitt, 2017). Starbucks provides a good example. Leveraging it's over 34 million fans on Facebook, Starbucks in mid-2017 posted on Facebook one-dollar offers for iced coffee, iced tea, or other frosty beverages. They did not tweet the offer to their nearly four million Twitter followers. From tapping its significantly larger Facebook following, analytics predicted higher visibility, increased incentive motivation, and a larger brand following on the single, but more highly-subscribed social media platform (Hemley, 2013). Thus, Starbucks successfully leveraged SM knowledge of a specific strategic strength (the characteristics of select, numerous Facebook followers), targeting prospective product customers to achieve a tactical benefit.

Strategic and tactical decisions are strongly interrelated through SM data. Because these data indicate how consumers, platforms, and firms use and form relationships with SM, they are

essential in creating SM-based strategies for ongoing functions that include product and service development, marketing, pricing, partnerships, and customer acquisition (Aral et al., 2013). With over 100 million Facebook fans and 35 million Twitter followers, SM is a hugely important strategic and tactical data source for the Coca-Cola Company (DeMers, 2014). Coca-Cola closely tracks how its more than 500 products in over 200 countries are represented across SM. SM data analytics provides insight into who the customers are, where they are located, and what situations prompt them to discuss Coca-Cola products. As a result, the company in 2015 was able to calculate that its products were mentioned somewhere in the world an average once every two seconds. It determined that product references, images, and other indicators—as well as those of its competitors—offer a tactical advantage in funneling advertisements to potential customers. According to the company, Ads targeted in this way have a four times greater chance of being clicked on than other methods of targeted advertising, thus providing a strategic path to superior company growth (DeMers, 2014).

4.1.2.5.2 Increasing Revenue with SM Data

Facebook, Twitter, and other social media platforms generate enormous revenues by selling vast amounts of advertising. For example, in 2015, world businesses spent a total of $24 billion on SM advertising (Media Buying, 2015), where 38% of businesses applied more than 20% of their total advertising budgets on SM channels (Smith, 2017). Nine out of ten of U.S. companies are now active on social networks (Holmes, 2015); all Fortune 500 companies engage with their customers on Facebook, and 83% have followers on Twitter (Smith, 2017). Facebook is the predominant vendor of SM advertising, increasing ad revenues from $30 billion to $35 billion in

2015 and 2016, respectively (PEW, 2017). Innovative companies leverage SM data to increase revenues and net income (Walker, 2016).

Netflix is known for its ability to deftly collect, manipulate, and extract directed SM content towards keeping its nearly 90 million subscribers and enticing new ones to register. Netflix uses SM data analytics to recognize subscriber wants and then target individual viewers with customized products and messages, leading to increased revenues and clientele (Fiegerman, 2017). Insights gleaned from SM data strategically influenced the company to produce Netflix originals—movies and series products which have been competitive with those of traditional TV networks and premium cable channels.

Implementing this strategy, Netflix cites its original "Stranger Things" as engaging viewers, strengthening loyalty, and inducing friends to watch Netflix-produced offerings. The firm encouraged virtual collaboration with episodes of "Black Mirror," where it introduced a fictional technology called Netflix Vista. It also launched an engagement app in the original series "Luke's Diner," which allows users to rate each other based on their social interactions. Using an online Snapcode to unlock a special Snapchat filter, more than 500,000 unique fans viewed the branded filter more than 880,000 times (Poggi, 2016).

Thus, Netflix asserts that it owes much of its marketing success to SM data analytics, which directed it toward producing originals. In 2017, the company plans to release over 1,000 hours of originals to satisfy subscribers and attract prospective buyers, with the goal that of increasing original content to 50% of its material. Netflix claims success, noting its third-quarter streaming revenue exceeded $2 billion for the first time. It added 370,000 new subscribers and raised its number of total streaming subscribers to nearly 90 million worldwide (Poggi, 2016). Innovative companies also leverage SM data to increase profitability and valuation (Walker, 2016).

4.1.2.5.3 Increasing Profitability and Annual Value of SM Data

While SM data can increase revenue and individual firm profitability, it also increases total firm worth. One measure is its aggregate potential annual value, which is the possible proceeds that can be generated from an entity, such as an investment, business, product, property, or other resources (Fama et al, 1998). For example, research indicates that SM data generates $300 billion potential annual value to U.S. healthcare, contributes €250 billion to European government services, and promotes a $600 billion in potential annual consumer surplus from synergizing personal location and experience data (Chen et al., 2014).

For example, United Airlines contributes to net income by using SM data analytics to improve the customer experience. The company employs its "collect, detect, act" system that examines 150 variables in each customer profile. The system compares previous customer purchases to customer priorities, generating a customized offer, reportedly increasing the company's year-over-year revenue more than 15% on $4.8 Billion in net income (United Airlines, 2016).

Contributing to annual value, two-thirds of the approximately $300 billion value in the healthcare sector would be generated by lowering healthcare expenditure. One way is emphasizing the delivery of personalized medicine (generating individualized diagnoses and prescribing treatments tuned to each patient's risk profile) delivers effective patient care and increases patient empathy. Another way is deploying clinical decision support systems that automate the analysis of x-rays, computed tomography (CT) scan images, and magnetic resonance imaging (MRI); the systems also employ analytics to mine the contemporary medical libraries to customize treatments to individual patients. Finally, the Centers for Medicare and Medicaid Services (CMS) has

demonstrated that SM data analytics tools are effective for fraud detection and prevention (Roski, et al, 2014).

4.1.2.5.4 Achieving Competitive Advantage with SM Data

When SM data leads to superior company performance over its competitors, it provides an opportunity for competitive advantage. Deft use of SM data analytics can produce unique, powerful, and high-speed sources of analytical content (Cheong, 2011; Schrecl & Keirn, 2013). As noted previously, imaginative businesses apply this content to refocus their competitive strengths to better meet customer expectations and improve profitability, thus seeking to ultimately attain competitive advantage (Manyika et al., 2011).

For example, a case study that focuses on customer expectations and profitability mined user-generated SM data from three major pizza chains extracted from Twitter and Facebook. The competitive analysis indicated that these pizza chains successfully engaged their customers through SM platforms by not only promoting their services but also bonding with their customers. Findings suggest that SM data play an important role in competitively sustaining a positive relationship with customers, which can lead to competitive advantage (He et al., 2013).

4.1.2.6 The Challenges of SM Data

There are several challenges involved with SM Big Data, such as capturing, storing, searching, analyzing, and virtualizing (Zaslavsky et al., 2013). Researches have classified the Big Data challenges into numerous categories. A study classified the challenges into engineering challenges, which relate to data management activities performance, and the others are related to information extracting from massive volumes of unstructured data (Bizer et al., 2012). Another

study categorized the challenges into data management challenges, which can be demonstrated in the integration between the structured and unstructured data (Zaslavsky et al., 2013). Katal et al. labeled Big Data challenges and issues into six different classifications, privacy and security, data access and sharing of information, storage and processing issues, analytical challenges, skill requirement, and technical challenges (2013).

There are many analyzing approach to analyze unstructured data, and data mining is the most popular technique (Schreck &Keim, 2013). Because SM data mining based on a human scale (Scarf, 2012), analyzing unstructured data skills is one of the challenges. A panel discussion arranged by FleishmanHillard company addresses how data impacts our decision-making across all aspects of social media and states “Data Needs Humans as Much as Humans Need Data” (2016). SM data demand on experts to discover how valuable it is and how the business can benefit from. Therefore, It is required advanced skills (e.g. technical, research, analytical, interpretive and creativity skills), the right background, and new way of thinking to extract and transform the data into useful information or knowledge for future actions (Kaplan et al., 2010; Manovich, 2011; Rajaraman et al., 2012;Katal et al., 2013). However, there are still some businesses lacks of understanding of the process of mining social media data, (He et al., 2013).

4.1.2.7 Recommendations for Companies for Exploiting SM Data Analysis

Social media data benefits organizations of any size and in any industry – from startups and small businesses, to multinational corporations (Rasp, 2016). Kaplan et al. (2010) provide such a set of recommendations and advice for companies which decide to utilize social media, and split the advice into two sections: one about using media and the other about being social. However, this study provides recommendations for companies integrating and adopting social

media data to their system to improve decision-making related to products or service. Our paper focuses on three main keys to benefiting from social media data.

4.1.2.7.1 Define the Business Goal for Analysis of SM Data

First, business needs to define the problem that needs to be solved or improved and determine the goal in analyzing data. For example, if the goal is to examine the overall reaction about the product or services, the business needs to determine the of SM sentiment and know if statuses, tweets, impressions, or comments are negative or positive. It is useful to focus on specific words and hashtags related to the product or service instead of analyzing random social media data.

4.1.2.7.2 Understand the Data

Social media data contains knowledge for businesses to leverage for a competitive advantage. In order to take advantage of the data value, business needs to understand each type of data that is produced from each SM sites and study the data flow and investigate it for each SM site. First, the business needs to define the users (e.g., demographic) and be aware that each social media site usually attracts a certain group of people (Kaplan et al., 2010). Second, the business needs to identify opportunities with data and discover the pattern, for example, how consumers speak about purchasing product (e.g., people behavior). There are numerous data analysis approaches and methods. For unstructured data analysis approaches and methods, Business Intelligence, Cluster Analysis, Data Mining, and Predictive Modeling are appropriate data analysis techniques. On the other hand, Crowd Sourcing, Textual Analysis, Sentiment Analysis, Network

Analysis and Analytics 3.0 era tools are appropriate analysis methods for unstructured and semi-structured data (Gandomi & Haider, 2015).

Data mining technology is used to explore patterns in large amounts of data. It includes statistical analysis, decision trees, and visual graphics to help in analyzing information for business decisions (Shaw et al., 2001). Text mining is one field of data mining that allows users to analyze large amounts of text data from different sources and to extract the topics and terms from the information (Hearst, 2003). Sentiment analysis is used for analysis of data from social media sites such as Twitter and is a natural language processing approach used to identify and extract subjective information (Kouloumpis et al., 2011). A study illustrates how social media data can be transformed into knowledge through text mining to analyze unstructured text content on the Facebook and Twitter sites of the three largest pizza chains: Pizza Hut, Domino's Pizza and Papa John's Pizza. The results reveal the value of social media competitive analysis and the power of text mining as an effective technique to extract business value from the vast amount of available social media data. Recommendations are also provided to help companies develop their social media competitive analysis strategy (He et al., 2013). Third, businesses need to recognize influencers.

4.1.2.7.3 Adopt Technology that Aligns with Your Goals

Capturing, storing, searching, analyzing, and virtualizing SM Big Data require certain technology and techniques. There are tools designed to manage and analyze unstructured data, such as those based on Hadoop, a software platform that can store huge files and process the information. Several researchers have evaluated Big Data application technology performance using software such as Hadoop and MapReduce. Hadoop is an Apache open-source software

framework for distributed storage and for processing large data sets on computer clusters built from commodity hardware. Hadoop consists of two parts: the storage part, called Hadoop Distributed File System (HDFS), and the processing part, called MapReduce (Hadoop, 2009). MapReduce is a parallel processing model for processing and generating large data sets (Dean, & Ghemawat, 2008). A study has examined the performance of MapReduce in placing the data across nodes. The researchers conclude that the strategy of rebalancing data across nodes before execution of the data-intensive application in a heterogeneous Hadoop cluster will always enhance the performance of MapReduce (Xie, Ding, Majors, & Qin, 2010). Another study on the relationship between Big Data and performance measurement systems offers insights into how measurement systems' performance can be influenced by Big Data analytics and strategy challenges and provides support for continuous improvement (Mello, Leite, & Martins, 2014).

4.1.2.7.4 Hire Competent Big Data Professionals

Because unstructured data analysis demands human skills, businesses need to hire and train people who can make sense of social-media data and use the technology and techniques. Business experts suggest that transforming SM data into knowledge requires professional people that have advanced skills so that they can deal with the huge volume and velocity of data (Scarf, 2012). The skills required to deal with SM Big Data are the technical, research, analytical, interpretive and creative skills (Katal, 2013) of people with a background of computer science, statistics, math (Manovich, 2011). Moreover, the Universities need to introduce curriculum on Big Data to produce skilled employees with this expertise (Katal, 2013).

4.2 Essay 2 Social Media's Influence on Millennials: A Case Study of the 2016 American Presidential Election

This research aims at analyzing social media influence on millennials' decisions regarding the 2016 American presidential election candidates. This work will contribute to a better understanding of the influence of social media during the 2016 American presidential election. The research integrates social influence theory, TAM2, and media richness theory and develops a model to examine individuals' attitudes toward participating in social media (SM) discussions that might influence their decision in choosing between the two presidential election candidates, Donald Trump and Hilary Clinton. The findings show that three factors impact the individuals' attitudes about participating in SM discussions: perceived usefulness, perceived ease of use, and SM community identification. Furthermore, perceived social influence has a significant positive relationship with SM community identification.

4.2.1 Introduction

The social media are a powerful source of data, news, and opinion expression (Valenzuela, 2013). Further, participating in social media discussions creates a communication channel that influences millennials' attitudes and decisions (e.g., about politics) (Goodrich et al., 2014). The Millennial Generation, or Generation Y, is the native digital generation (Prensky, 2001), and it has become the largest generation in the United States, with a current population of around 79 million (Knoema, 2016). Millennial Generation members were born in a social media environment (Bolton et al., 2013). More than 86% of the people who use at least one social media site are Millennials (Pew-Research, 2016). This number is a testament to the significant role of social media in the lives of millennials and to the extent to which they are involved in their daily lives (Ellison, 2007)

and have shaped their attitudes and decisions based on their discussions with their peers (Rohampton, 2017).

Therefore, it is important to include social media influence in any discussion of the social influence on an individual's decision in a particular event. The social media sites are an important source of social influence on individuals (Asur et al., 2010; Cheung et al., 2011). Social influence is defined as "change in an individual's thoughts, feelings, attitudes, or behaviors that results from interaction with another individual or a group" (Rashotte, 2007). Social media (SM) influence is the change in an individual's thoughts, feelings, attitudes, or behaviors that results from using social media sites. Schmitz et al. describe the social influence of social media use as social information (1991). Social information on social media sites is the information that an individual shares about him/herself that reflect feelings, opinions, etc. (Campos, 2013). This research focuses on social media influence on millennials' attitudes and decisions that results from their participation in social media discussions.

An SM discussion platform is a tool that provides an easy way to share one's opinion on any topic that one wishes to discuss with others (Kim et al., 2013), and it is one source of social media data. The opportunity to participate in an SM discussion can attract a large audience to discuss and express opinions about any topic, such as a political event, particularly if participants belong to the same social group (Yeoman, 2017); such a platform produces a huge volume of data that can be of use in the prediction of future political events. In this study, the 2016 U.S. presidential election was selected as an important event influenced by SM because that election was different from previous elections. The difference was largely due to the increased role of the social media (SM) that were used to communicate candidates' campaign information. SM clearly played a role in the 2016 U.S presidential election, and that was not the first time that had

happened. Yet, the effect SM had on people was more intense this time around than ever before. The two main candidates, Donald Trump and Hillary Clinton, both used the power of SM to their advantage. In particular, the supporters, influencers, and surrogates of each of them shaped perceptions to serve the interests of their preferred candidate through social media (Michaels, 2017). The goal of this research is to investigate the social media influence which can help to explain the biases reflected in the millennials' decisions.

Prediction of the successful candidate in the 2016 U.S. presidential election using social media data analysis might have been possible to achieve. However, there was an underestimation of the role of social media's (SM) influence on the event, as well as flawed analysis of social media information, and these ultimately caused incorrect predictions of the election outcome. Thus, the victory of Donald Trump was a largely unexpected result. This was attributable to incorrect interpretation of social media's influence, which resulted in inaccurate predictions and faulty analysis outcomes.

This study seeks to better understand the impact of social media influence in decision making and because of its general applicability to many groups uses the 2016 U.S. presidential election as the venue for testing such influence. Examination of the potential attitudes surrounding the use of SM discussion platforms and how it influences individual decisions is difficult because of the varied interests of different market segments. For example, some individuals rely on social media for information about restaurant decisions, banking service, or cars, but these interests do not cross across a broad range of different individuals. However, interest in the U.S. presidential election is often broad and cuts across many different groups. As a result, the broad interest in the presidential election presents a rich opportunity to study how social media influences decisions. Despite the increasing popularity of social media as the major source of data, little research has

been carried out regarding how social media's influence contributes to the shaping of political decisions. While many studies have examined the role of social media use and explored the strong influence of using of social media sites on political activities and participations (Groshek et al., 2013; Dimitrova et al., 2014; Stephens, 2016), few studies have investigated the social media's influence on millennials' attitude to participate in SM discussion that impacts voting decisions.

Millennials are considered as the most significant user of social media (Sago, 2010), the objective of this study is to examine how information conveyed to Millennials through SM (i.e., on SM discussion platforms) affected their decisions as measured by how they chose between the two presidential candidates. This research also aims to fill a gap in Data Science, Information Systems, and Social Science literature by investigating the mediating role of a specific community in the association between perceived social influence and millennials' attitudes toward participating in SM discussion platforms and by exploring the impact of social media on decisions as measured by political decisions. Pursuant to the objectives of the study, the main research questions are:

- How did social media data shape the 2016 presidential election?
    - Are millennials influenced by people whom they follow on social media?
    - Did social media use shape millennials' voting decisions?
    - Did people's attitudes toward participating in social media discussions influence millennials' political decisions?

In the next section, we address the prior literature in the areas of SM influence with a focus on a political decision. Using this as a foundation, the research model and hypotheses, including the Social Influence theory, Technology Acceptance Model, and Media Richness theory constructs, are presented. The next section describes the research methodology including the sample and survey method, and the following section presents data analysis and results, followed

by discussions, implications, and contributions of the findings. The final two sections discuss research limitations and future research directions and present the research conclusion.

4.2.2 Literature Review/Theoretical Background

4.2.2.1 Social Media Influence and Millennials

The study of social media's influence on millennials has been active in several fields, particularly in communication, marketing, and political science. Millennials are referred to as the "Next Generation" and "Generation Y." They were born between 1982 and 2000 and represent a significant segment of the population (30 percent) (Yerbury, 2010). It is their early exposure to the internet that distinguishes them from other generational cohorts (Bolton et al., 2013). A study examines the levels of influence, both positive and negative, from the users' comments obtained via social media on older members of the Millennial Generation (Sago, 2010). Another study states that communication through social media helps people share information and resources, which can impact personal and managerial decision making. It also helps to alter individuals' opinions and to influence their choices by impacting decisions of consumers and business decisions of managers (Power et al., 2012). Moreover, a study by Pinto et al. (2015) addresses social media's influence on individual's decision process regarding buying tourism products and indicates that individuals are influenced by opinions, comments, reviews, and reports posted on social media sites. There are many factors that influence a millennial's attitude to participate in SM discussions and make decisions. This section describes social media use as affected by social and technical factors.

4.2.2.2 Social Influence Theory

Individuals use social media sites to interact with, share with, and obtain political information from others, which indicates that there might be a direct impact of social media use on millions of users and their friends and friends of friends, particularly on their political self-expression, information seeking, and voting behavior (West, 2013). Social influence theory can explain what happens when individuals change their attitude or behavior as a result of inspiration by other persons or groups (Kelman, 1961). The theory suggests that three social processes, compliance, identification, and internalization, affect individual attitudes in making voting decisions (Kelman, 1974). For example, individuals may react to their followers' opinions in discussion posts on social media sites. Also, individuals can identify their feelings about specific groups or communities that influence their voting decisions and adopt followers' opinions due to the similarity of their own values with those of their followers (Zhou, 2011; Dholakia et al., 2004; Venkatesh & Davis, 2000; Cheung et al., 2011).

4.2.2.3 The Influence of SM on Political Decisions

The usage of social media in the political field started in the 2008 election campaign. It played a significant role in observing and in affecting voters (Hesseldahl al., 2008; Marchese, 2008; Owen, 2008). During the 2008 election, young adults relied on social media more than on traditional media for sharing information, obtaining campaign news, and expressing political opinions about candidates more than members of other generations did (Kohut, 2008; Smith & Rainie, 2008). Other have found that it is effective for achieving political results and increasing social capital (Kim & Geidner, 2008; Utz, 2009; Valenzuela, Park, & Kee, 2009; Vitak et al., 2010). More studies on examining political social media use and the influence of social media on

participation in specific political activities such as joining Facebook groups and visiting candidate profiles indicate there is a relationship between the use of SM and political activity (Dimitrova et al., 2014). Other studies indicate that SM allow people psychologically to engage in political processes and that media may engage people in the political process because of the communication of other users (Groshek & Dimitrova, 2013; Stephens, 2016). Kohut (2008) concluded that as political actors made more use of SM for campaigning, young adults began to rely on online media more than traditional sources for information about political developments. Several observers of the media and politics have concluded that SM influenced young voters' thinking and behavior in the 2008 election (Keeter et al., 2008). Zhang et al.'s (2010) study addresses the influence on political attitudes and democratic participation and shows a significant increase in civic participation, but not political participation. (Zhang at. al, 2010). Another study identified a positive and significant predictor of people's social capital and civic and political participatory behaviors, online and offline (Zuñiga, 2012). Holt and Shehata (2013) investigated the effect of social media use for political purposes on younger or older people and discovered that there was a difference of social media usage between young people and older ones; young people pay attention more social media information than traditional media. A significant sign has been observed in the use of social media for political engagement among young people and the influence of the new popular digital media on long-standing patterns of political inequality (Xenos & Vromen, 2014). However, other researchers have concluded that the use of SM does not have a significant effect on behavior and thinking about politics (Ancu & Cozma, 2009; Gil de Zuñiga, Puig, & Rojas, 2009; Zhang, et al., 2010; Stieglitz, 2012; Dimitrova et al., 2014; Yamamoto, 2015; Zafer 2017). Also, although social media allow users to experience and obtain political content, the relationship between political self-efficacy and situational political involvement and social media political use

and influence on voter decisions during the local elections (2010/2011) is insignificant (Kushin & Yamamoto, 2010; Effing & Hillegersberg, 2011).

4.2.2.4 Technology Acceptance Model (TAM2), Social Influence, and SM

Researchers have emphasized the effects of social influence on user acceptance of an information technology to understanding the role of social influence in the TAM (Davis et al., 1989; Malhotra et al., 1999). Individuals use several social media sites such as Facebook, Snapchat, Twitter, Google Plus, or any new social media site to interact and to share and obtain political information. There are several factors that influence their decision about how and when they will use such a site (Venkatesh & Davis 2000).

4.2.2.5 Media Richness Theory

Media richness theory describes a communication medium's ability to reproduce the information sent over it; the theory is an extension of information processing theory (Daft & Lengel, 1986). Media classifications range from face-to-face interactions to numeric documents (Liu et al., 2009). Dennis & Valacich extended the original four dimensions of media richness theory to five dimensions: immediacy of feedback, parallelism, symbol variety, processability, and reversibility; the resulting approach is referred to as the theory of media synchronicity (1999). SM platforms are important channels that deliver information in different forms such as text, picture or video, etc. (Kaplan & Haenlein, 2010). Researchers indicate that the richness of media has significant positive impacts on decision quality when participants' task-relevant knowledge is high (Power et al., 2012).

4.2.3 Research Model and Hypotheses

This research proposes hypotheses about how participation in SM discussions affects individuals socially and influences their attitude in making voting decisions, in an effort to explain the unexpected outcome of the 2016 U.S. presidential election in the context of social influence theory showing that participating in social media sites influences individuals' attitudes in making voting decisions (Varnali et al., 2015). Based on previous work on social influence, media richness, and SM influence, the authors of this study define the key constructs, develop hypotheses, and put forward a conceptual framework, as shown in Figure 4.1.

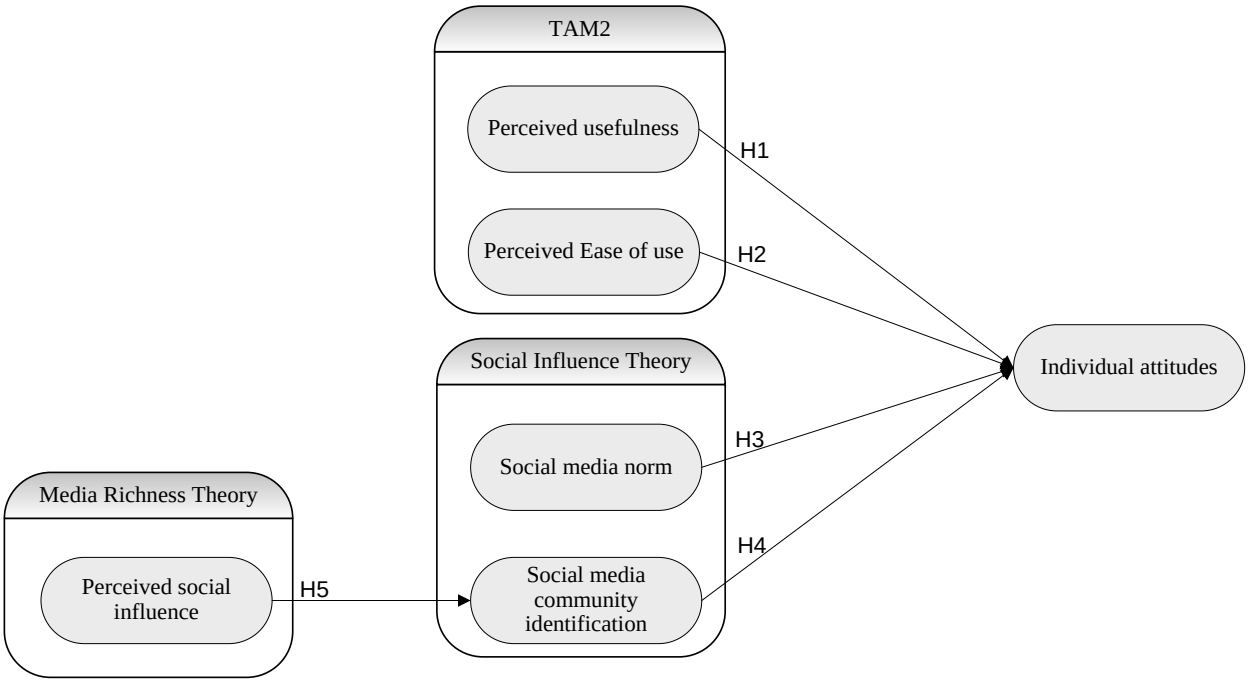

Figure 4.1: Conceptual framework

4.2.3.1 Perceived Usefulness and Perceived Ease of Use

These two constructs are drawn from TAM2. This study adopted perceived usefulness and perceived ease of use about social media from Porter et al., (2006) and from Hsu & Lin (2008). Perceived usefulness is the degree to which an individual believes that using a SM platform

influences his or her performance in making decisions. Perceived ease of use is the degree to which an individual believes that using an SM platform is free of effort. The more that an individual perceives the SM platform as useful and easy to use, the more favorable that individual's attitude toward the use of SM discussion platforms will be (Porter et al., 2006). Thus, we propose the following hypotheses:

H1: There will be a significant positive relationship between perceived usefulness and individual attitudes.

H2: There will be a significant positive relationship between perceived ease of use and individual attitudes.

4.2.3.2 Social Media Norm (SN) and Social Media Community Identification (CI)

The social influences theory provides theoretical bases for a relationship between social norms and individuals' attitudes (Venkatesh et al., 2003). This study adopts the social norm concept from Hsu & Lin (2008). It defines social media norms as related to the level at which an individual recognizes that his/her choices and attitudes are endorsed by others, through participation in SM discussions. Empirical research indicates that individual attitudes about participation in SM discussions influence their voting decisions (Silver et al., 1986). An individual's attitude toward participating in SM discussions with others that share the same norm can influence the individual's voting decisions. Also, this study adopts the community identification in social media concept from Hsu & Lin (2008). Community identification in social media sites leads to a sense of belonging to a particular group among members of a SM discussion platform. An individual's attitude toward participating in SM discussions with the group to which the individual belongs can influence the individual's voting decisions. Thus, we propose the following hypotheses:

H3: There is a significant positive relationship between social media norms and individual attitudes.

H4: There is a significant positive relationship between community identification and individual attitudes.

4.2.3.3 Individual Attitudes

This study adopts the individual attitudes concept from Hsu et. al., (2008). Individual attitude as used in this study is the preference for participating in SM discussions, which may have influenced the individuals' decisions to vote for a 2016 U.S. president candidate. Use of SM sites in general influenced people's attitudes socially and politically (Moy et al., 2005; Shah et al., 2001; Wellman et al., 2001).

4.2.3.4 Perceived Social Influence

Perceived social influence is a construct drawn from media richness theory; it is the change in an individual's thoughts, feelings, attitudes, or behaviors that results from interaction with another individual or a group (Rashotte, 2007). For this study, we adopt the perceived social influence concept from Carlson et al., (1999). An individual can be influenced socially by his/her group's posts such as discussions, image, videos, etc., on SM sites. We propose the following hypothesis:

H5: There is a significant positive relationship between perceived social influence and social media community identification.

4.2.4 Methodology

After the 2016 election, the researchers for the current study conducted an online survey of college students at a large public university in the southwestern U.S. The focus of this study is

on millennials' use of social media (SM); specifically, undergraduate college students who are enrolled in business classes and political classes are the population for this study. Research indicates that college students rely much more than older adults on SM as a source of political campaign news (Pew Research Center, 2016).

The authors developed a survey instrument by adapting established measures from prior studies. Measures of social media norms and social media community identification were all modeled and adopted from previous research and contextualized for this research (Hsu & Lin, 2008). TAM2 constructs (usefulness and ease of use) were operationalized and measured using items adapted from other research and using author-developed scales (Venkatesh & Davis, 2000). Measures for perceived social influence were adopted from Carlson et al., (1999). The measure of Attitude was adopted from Webster and Trevino's (1995) study. For a complete list of measures, see Appendix A. Seven-point Likert scales were used to capture student responses.

After receiving approval from the University's Institutional Review Board, the researchers approached instructors, who posted a link to the survey on course websites; the survey was administered online. All students were offered extra course credit to encourage participation. Students in a total of 10 classes were asked to complete the survey. The authors received a total of 1,101 responses, including 195 from international non-voters. After cleaning the data to eliminate the unusable responses, including those that indicated a lack of variance (i.e., from respondents selecting all 1's or all 7's) and incomplete surveys, 450 usable responses remained for further analysis, resulting in a 40% response rate. The sample achieved one of the goals of the research, that of targeting younger participants; 68% of the respondents were under the age of 21. Most participating students were male (52%). Most students (55%) had voted for Hillary Clinton, and

27% of the students had voted for Donald Trump. Complete survey demographics are provided in Table 4.2.

Table 4.2: Respondent demographics

Gender			Age			Voting		
Male	232	51.56%	18-21	308	68.44%	Donald Trump	123.00	27.33%
Female	216	48.00%	22-25	85	18.89%	Hillary Clinton	248.00	55.11%
Others	2	0.44%	26-29	27	6.00%	Gary Johnson	32.00	7.11%
Academic Statues			30-33	13	2.89%	Jill Stein	11.00	2.44%
Freshman	118	26.22%	34+	17	3.78%	Other	36.00	8.00%
Sophomore	106	23.56%	Hillary Clinton voters			Donald Trump voters		
Junior	129	28.67%						
Senior	72	16.00%	Female	141	31.33%	Female	42	9.33%
Graduate	25	5.56%	Male	106	23.56%	Male	80	17.78%

4.2.5 Data Analysis and Results

We tested the model and measured the reliability and validity of the model constructs. We used SmartPLS version 2.0 to measure the overall fit of the complete structural model and to calculate beta coefficients that explain the construct relationships and the average variance extracted (AVE). The results indicate that the model instrument satisfies reliability because Cronbach's α for each construct exceeds the minimum score of 0.7 for exploratory research (Nunnally 1978; Nunnally & Bernstein, 1994). In addition, we used SmartPLS to run Confirmatory Factor Analysis (CFA). All five constructs exhibited factor loadings exceeding .7 on the expected factor (Hair et al., 2010), and the values of the variance of the latent variable explained by the indicators, which is presented on AVE, was in the range 0.6 to 0.7 which exceeds the minimum value of 0.50 (Chin, 1998; Henseler, Ringle, & Sinkovics, 2009). Table 4.3 presents a summary of the reliability analysis for each latent variable.

Table 4.3: Measurement model summary.

							Factor Correlations					
Item	Mean	Std. Deviation	Factor Loading	Cronbach's α	Composite Reliability	AVE	SMN	SMCI	PSI	PEU	PU	Att
SMN1	3.317	1.595	0.843	0.773	0.870	0.642	0.829					
SMN2	4.430	1.608	0.897									
SMN3	4.268	1.576	0.746									
SMCI1	4.254	1.614	0.827	0.749	0.841	0.697		0.753				
SMCI2	3.212	1.518	0.794									
SMCI3	2.922	1.179	0.751									
SMCI4	2.672	1.508	0.640									
PSI1	3.059	1.187	0.797	0.855	0.902	0.741			0.834			
PSI2	2.762	1.095	0.832									
PSI3	3.060	1.391	0.848									
PSI4	3.102	1.347	0.860									
PEU1	4.916	1.345	0.837	0.817	0.872	0.733				0.792		
PEU2	1.269	1.418	0.676									
PEU3	2.808	1.133	0.824									
PEU4	3.055	1.225	0.833									
PU1	2.500	1.585	0.904	0.904	0.930	0.609					0.881	
PU2	2.394	1.580	0.905									
PU3	2.590	1.597	0.919									
PU4	1.981	1.735	0.795									
At1	2.209	1.741	0.938	0.932	0.957	0.631						0.938
At2	3.584	1.632	0.944									
At3	3.701	1.701	0.933									

SMN = Social media norm, SMCI = Social media community identification, PSI = Perceived social influence, PEU = Perceived Ease of use, PU = Perceived usefulness, Att= Individual attitudes.

Furthermore, we analyzed the structural model by calculating t-tests for each path and R square values using 5,000 bootstraps as recommended by Hair et al. (2014, 132) on the 450 sample data points using SmartPLS (Ringle, Wende, & Will, 2005). Four of the hypothesized paths were statistically significant at the .05 level, while the fifth path (SM norm) was insignificant. The constructs perceived usefulness, perceived ease of use, and SM community identification all had significant positive relationships with individuals' attitudes. Perceived social influence had a significant positive relationship with SM community identification. Figure 4.2 shows a summary of the structural model.

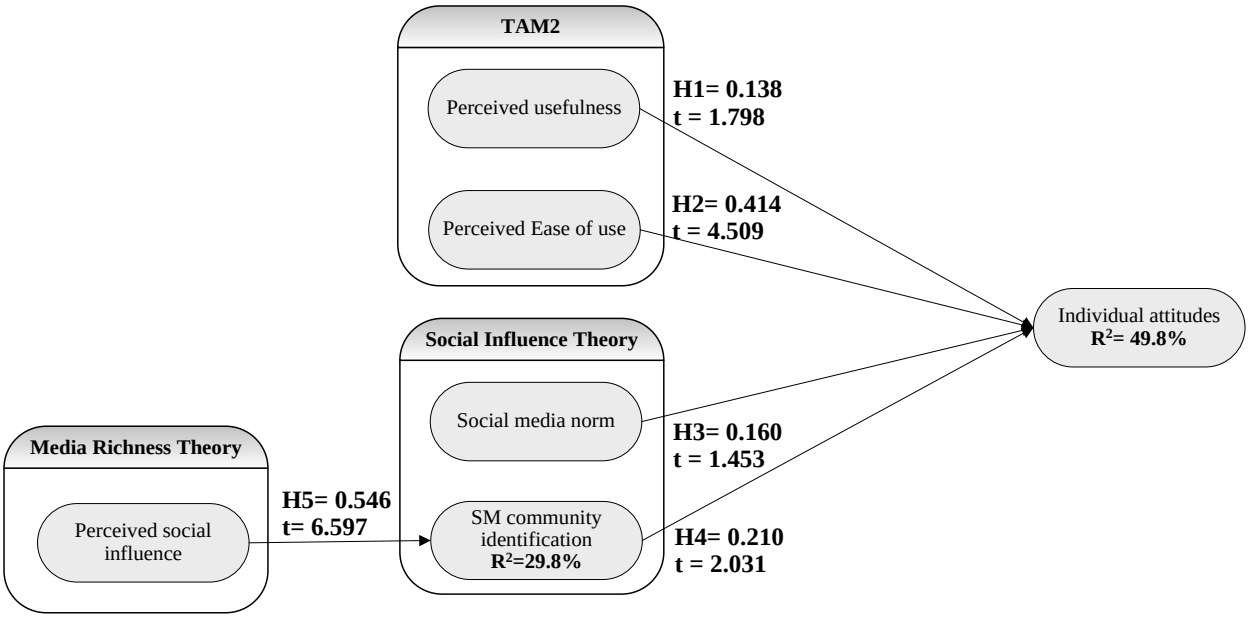

Figure 4.2: PLS structural equation path results. Path significant at $P < 0.05$ level

Finally, we used SPSS to calculate the chi-square tests for independence (McHugh, 2013), comparing two variables: Candidates (e.g., Donald Trump, Hillary Clinton) and Individuals' Attitudes, presented in a contingency table (Table 4.4) to find out if there was a relationship between individuals' attitudes toward participating in SM discussion platforms and voting for certain 2016 U.S. presidential candidates. There was a significant relationship between the individuals' attitudes and voting for a specific presidential candidate. Table 4.5 shows a summary of the chi-square test results.

Table 4.4: Candidates * individuals' attitudes categories cross tabulation

Candidates		Disagree	Individuals' Attitudes Categories	Agree	Total
Donald Trump	Count	40	11	73	124
	Expected Count	28.5	8.8	87.1	124.0
	% within Candidates	32.3%	8.9%	58.9%	100.0%
	% within Individuals' Attitudes Categories	47.1%	44.0%	28.1%	33.5%

	% of Total	10.8%	3.0%	19.7%	33.5%
Hillary Clinton	Count	45	14	187	246
	Expected Count	56.5	16.6	172.9	246.0
	% within Candidates	18.3%	5.70%	76.0%	100.0 %
	% within Individuals' Attitudes Categories	52.9%	56.00%	71.9%	66.5%
	% of Total	12.2%	3.80%	50.5%	66.5%
Total	Count	85	25.00%	260	370
	Expected Count	85	25.00%	260.0	370.0
	% within Candidates	23.0%	6.80%	70.3%	100.0 %
	% within Individuals' Attitudes Categories	100.0%	100.00%	100.0%	100.0 %
	% of Total	23.0%	6.80%	70.3%	100.0 %

Table 4.5: Chi-square tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	11.682[a]	2	0.003
Likelihood Ratio	11.389	2	0.003
Linear-by-Linear Association	11.228	1	0.001
N of Valid Cases	370		

a.No cells (0.0%) have expected count less than 5. The minimum expected count is 8.38.

4.2.6 Discussion

This research examines the contribution of individuals' attitudes toward participating in SM discussion platforms to their decisions to vote for specific presidential candidates. The results from the structural model show that individuals' attitudes toward using SM sites and their platforms were influenced by certain groups who shared similar interests because individuals perceived the SM site platforms as useful, easy to use, and favorable. Three characteristics,

perceived usefulness, perceived ease of use, and SM community identification, influenced individuals' attitudes toward participating in SM discussion platforms and explained 49.8% of the variance. Interestingly, perceived social influence strongly influences the SM community, explaining 29.8% of the variance as shown in Figure 4.2. In addition, the chi-square test result indicates that there was a relationship between Individuals' Attitudes toward participating in SM discussion platforms and their choice between the two presidential candidates, explained in Table 4.2, allowing us to build the structural model.

4.2.7 Conclusions

This research develops an SM influence model by integrating social influence theory, TAM2, and media richness theory and provides insights into the influence of SM on voting decisions. The results indicate that perceived social influence is a major factor because it significantly influences SM community identification, while SM community identification strongly influences individuals' attitudes. A major contribution of this study is its indication that SM influences individuals' decisions in choosing between the two presidential candidates. Though the data in this study were collected from students who had voted for the 2016 U.S. presidential candidates, the authors believe the model can be generalized to other examples, or to other types of events where SM sites are a major source of data.

4.3 Essay 3 Big Data Professions

This research defines the major job descriptions for careers in the new Big Data profession. It to describe the Big Data professional profile as reflected by the demand side, and explains the differences and commonalities between company-posted job requirements for Data analytics,

Business analytics, and Data Scientists jobs. The main aim for this work is to clarify of the skill requirements for Big Data professionals for the joint benefit of the job market where they will be employed and of academia, where such professionals will be prepared in data science programs, to aid in the entire process of preparing and recruiting for Big Data positions. We compare the five fields against the unified back drop of common semantic dimensions, and we examine the recent dynamics. As a result, that there is a need for communication skills and statistical thinking in the context of working with data.

4.3.1 Introduction

In the past few years, with the emergence of the Big Data phenomenon, organizations have recognized the opportunities for competitive advantage and the potential business value they can gain from incorporating Big Data capabilities into their organizational structure. As a consequence, organizations have focused their attention on acquiring tools and skills involved in Big Data analytics. However, as they engage in piloting or deploying Big Data initiatives, a gap in technical and analytic skills in the workforce is one of their top challenges (Schroeck et al., 2012). Adapting their talent acquisition process to this development, they started recruiting for positions such as data scientists, statisticians, business analytics professionals, etc. (Provost & Fawcett, 2013). The presence of so many overlapping positions confuses the workplace (Russom, 2013), which can impact business effectiveness (De Mauro et al., 2016). Ambiguity in describing jobs related to Big Data is a multidisciplinary problem involving job functions such as information technology, software development, database administration, statistical analysis, predictive modeling, etc. According to McKinsey Global Institute Report 2011, the USA could face a shortage of 140,000 to 190,000 people with deep analytical skills by 2018 (Manyika, 2011).

On the academic side, data science has emerged as the field that can address the industry needs for Big Data skills. Researchers have defined data science as "the computational aspects of carrying out a complete data analysis, including acquisition, management, and analysis of data" (Johnstone and Roberts, 2014). Many recent studies have recognized the job market as a major force that shapes the data science curriculum (e.g., Hardin et al. 2015, etc.) However, arguments on the exact set of skills that are appreciated in the workplace are subjective, based on anecdotal evidence and personal observations. Even industry reports are quite often survey-based, aggregating subjective opinions, experiences, and observations of the participants. Rather than studying the job market directly, such indirect views provide their own interpretation of what the job market needs. In the present study, we address this gap by collecting and analyzing job descriptions in five professions that relate to the overlapping areas of Big Data and data science: Statistical Analysts (SA), Big Data Analytics Professionals (BDA), Data Scientists (DS), Data Analysts (DA), and Business Analytics Professionals (BA). Our research aims at describing the Big Data professional, on the demand side, to the data science curriculum developer, (e.g., a program director), on the supply side.

The goal of the current research is to clarify skill requirements for Big Data professionals for the joint benefit of the job market where they will be employed, and academia, where such professionals will be prepared in data science programs. We pursue our goal by examining differences and commonalities among company-posted job requirements for the five professions listed in the previous paragraph. By accomplishing this goal, we hope to contribute to the improvement of the entire process of preparing and recruiting for Big Data positions, and to address the need for clarification of such overlapping subjects not only in industry but also in the academic environment (Mauro et al., 2016). To achieve this goal, this research evaluates several

industry job descriptions for Big Data positions using automated content analysis. The questions we seek to answer are as follows:

RQ1: Within the Big Data domain, what are the differences and commonalities among job description elements for SA, BDA, DS, DA, and BA?

RQ2: What has been the dynamic behavior of these differences and commonalities in recent years?

4.3.2 Big Data Professions

Several business reports have underscored the need for understanding the recent changes in the Big Data profession since it was recognized as early as the 1999 ISI meeting in Helsinki (Friedman 2001). Small firms frequently take advantage of their existing employees' skills and create value with Big Data project initiatives and by implementing Big Data applications (Davenport, 2014). On the other hand, research indicates that there is an overlap of skills when it comes to Big Data analysis (Hardin, Hoerl, & Horton, 2015). A report (Ahern & Keller, 2014) finds that Big Data professionals typically have an advanced degree in statistics, applied mathematics, operations research, or economics, and business (Cleary, &Woolford, 2010). The same report states that 86% of Big Data professionals have at least a master's degree, and 20% have a Ph.D. Nicholls (2001) recognizes the need to add IT-related courses that would enhance the training of the statistical analyst profession. In addition, Big Data professionals need to have experience in one or more advanced analytics tools and methodologies, such as data mining, modeling, and advanced coding and programming. The main professions in the field of Big Data are listed below.

4.3.2.1 Statistician/Statistical Analyst

Statisticians have recently found themselves in one of the hottest fields (Chamandy,

Muralidharan, & Wager, 2015). In 2017, U.S. News ranked the profession of statistician #1 in Best STEM Jobs, #1 in Best Business Jobs, and #4 in the overall category of The 100 Best Jobs (U.S. News 2017). Statisticians or statistical analysts can produce estimates, test hypotheses, design, analyze, and interpret experiments, prepare data for further analysis, and build statistical models that support business decisions. However, these skills are not enough to help them face the challenges of Big Data. In order for statisticians to deal with massive data from different sources and maintain Big Data, they need to have advanced degrees (e.g., Ph.D.) in mathematics, or related fields, and specialize in specific types of data, such as business, marketing, health and medicine, education, or economics (Varon, 2012; Hall, 2012). The society recognized the importance of using statistical tools and statistical analysis in order to support decision-making, but statisticians themselves were going through an identity crisis (Mason 2004). Moreover, with the developments and the current high demand for statisticians, Chamandy et al. (2015) emphasize the importance of computing and manipulation skills and propose the exposure of statistics students to tools such as MapReduce so that they can be prepared for the Big Data world.

4.3.2.2 Big Data Analytics Professional

With the emergence of the Big Data phenomenon in the scientific and business world, there is a demand for dealing with the Big Data dimensions of volume, velocity, variety, and veracity (Schroeck et al., 2012). There are storage needs that require clusters of servers and distributed file system architectures, the need to analyze streaming data without interrupting the stream, the need for parallel processing analytics, the need to analyze images, video, and unstructured text, and the need to manage “dirty” data with missing values that come from unreliable sources and have formatting errors. The duties of a Big Data analyst typically include inspecting, cleaning,

transforming, and modeling Big Data using analytical techniques and tools, such as correlations, cluster analysis, filtering, etc. to discover knowledge and support decision-making (Bi et al., 2014). The literature suggests that academic Big Data analysis courses should focus on multiple skills such as practical computational skills, data visualization, and statistical skills and have a broad understanding and experience with real-time analytics, business intelligent platforms, and several programming languages and software packages such as Tableau, SQL databases, R, Java, MatLab or SPSS (Hardin et al, 2015; Rijmenam, 2016).

4.3.2.3 Data Scientist

Hardin et al. (2015) describe the origin of data science as a mixture of the statistics and computing professions. Data science is typically understood as a combination of statistics, business intelligence, sociology, computer science, and communication (Stadd, 2014). Data science as a concept was proposed in 1999 (Cleveland 2001). Data science has no universal definition, but it is a discipline that involves discovering, extracting and analyzing data that can be used for decision-making and knowledgeable prediction (Ahern & Keller, 2014; DeVeaux et al., 2017). The National Science Foundation (NSF) has defined data science as "the science of planning for, acquisition, management, analysis of, and inference from data" (NSF 2014, p. 4). The importance of Data Science was recognized in the 2014 ASA report on the undergraduate curriculum in statistical science (ASA 2014). A more comprehensive list of tools and skills that data scientists are expected to master, would include such diverse items as knowledge of data management and storage tools such as SQL, modern computing and manipulation tools that can merge, aggregate, and process iteratively, familiarity with data graphics and elements of visual perception, confidence intervals via the bootstrap, simulation, regression, variable selection, data mining/machine learning,

classification, cross-validation, text mining, mapping, regular expressions, network science, and MapReduce, as well as many other additional topics (Hardin et al, 2015).

4.3.2.4 Data Analyst/Data Analytics Professional

Data analysts are hard to distinguish from data scientists. Their academic preparation starts with a degree in statistics, mathematics, computer science, management science, biological sciences, economics, information management, or business information systems. Data analysts need to have analytical skills, communication skills, technical skills, familiarity with relational databases, knowledge of data modeling, and experience in data analysis tools (Wladawsky-Berger, 2014; Stadd, 2014). At the work place, data analysts are expected to not only analyze data and gain insight but also to use their communication skills and perform visual, written, and verbal delivery of relevant information to intended audiences. Data analysts can work in business intelligence, data assurance, data quality, finance, higher education, marketing, and sales (Welch, 2017).

4.3.2.5 Business Analytics Professional

Business analytics professional, not to be confused with the much older term business analysts (who typically investigate current business processes and design technology solutions), is a relatively recent job description term. The online Gartner IT Glossary defines business analytics as a collection of "solutions used to build analysis models and simulations to create scenarios, understand realities and predict future states" (Gartner, 2017). Essentially, business analytics is the integration of business and data science (Russom, 2013). Since the scope of BA includes what was previously known as Business Intelligence (BI), a recent trend is to merge BI and BA through the term Business Intelligence & Analytics (BI&A). Business analytics professionals develop new

business insights and provide recommendations based on analysis of data. They need to have an IT background and statistical knowledge, with some added business experience. They also need to be familiar with data mining, predictive analytics, applied analytics, and statistics (Sallam et al., 2017).

4.3.3 Big Data Professions Responsibilities

Organizations are able to manage Big Data if they set clear job responsibilities and recruit the right people who possess the right skills. However, the creation of a successful Big Data team requires consideration of multiple aspects, including the right people, who assume the right roles in the way they participate in the data and data-driven solutions lifecycles, and the right service-oriented culture (Ariker, McGuire, & Perry, 2013). Based on our own observations of job postings at Monster.com, Big Data job responsibilities seem to center on several areas such as technical knowledge, business understanding, statistical analysis skills, data manipulation experience, etc. In addition, Big Data jobs seem to require familiarity with a number of advanced analytic methodologies, such as data mining, modeling, and advanced coding, and tools, such as SAS, R, Hadoop, and SQL, (Ahern & Keller, 2014).

Table 4.6 lists some example job responsibilities for the five professions included in our study, based on selected Monster.com website postings. Based on these examples, work places expect statisticians to function within a team, perform statistical analysis, produce information that match the business environment, and communicate the results to nontechnical individuals. Data scientists are expected to build statistical models and incorporate them within an IT environment that deploys and maintains multiple systems, platforms, and databases. Big Data analytics professionals are expected to implement analytics techniques to data from different sources,

support data management, and data quality processes, communicate with teams of developers, and provide business recommendations. Data analysts are expected to work with data, analyze the data using different analytics strategies and tools, present results to audiences that include executives, and possess teamwork and project management skills. Finally, business analytics professionals are expected to implement data-driven strategies, develop databases, collect and analyze complex data, and provide information solutions to various departments within an organization.

Table 4.6: Big Data profession responsibilities

	2015	2016	2017
SA	Perform data analysis using primarily the SAS programming language. Interact with Statisticians and another team, perform analysis and generate outputs.	Utilize data mining and statistical techniques developing analytic insights, recommendations. Communicate technical subject matter to individuals from various backgrounds	Provide assistance in data quality, imputation, and other statistical analysis. Provide technical support to design and develop projects.
DS	Create models and frameworks to better understand and predict user behavior. Assist in the development of analytics tools and systems to help build out the analytics platform.	Conduct multi-channel deep analysis for online and offline contexts sources. Use statistical tools and techniques to extract and analyze trends from the customer database warehouse.	Design market response analytics models and approaches. Know and understand our data elements and architecture deeply and comprehensively.
BD	Apply semantic correlation, ontology and text analytics techniques to analyze unstructured data and identify insights. Participate in all aspects of software lifecycle: analysis, design, development, unit testing, production deployment, and support.	Apply appropriate analytic and statistical methodologies to develop technology strategy to support data management, data quality. Develop recommendations for business, and IT teams to improve data quality and management and inventory.	Drive the analytics strategy and execution and manage products through various launch phases. Interact with partners, customers and prospects to define market requirements.

DA	Utilize mined data, apply statistical modeling, develop summarized reports.	Demonstrate strategic technical planning skills and leadership to ensure successful completion.	Perform database queries to analyze and validate data.
	Play leadership role in driving analytic strategies between our business customers and IT services.	Document data entities, attributes, business access rules, and data volumes and retention, and design the logical database.	Perform detailed data analysis to support changes to reports and/or processes.
BA	Help support the identification and implementation new data-driven strategies and processes for the organization.	Develop specific databases for collection, tracking, and reporting to analyze, review, forecast, and trend complex data.	Apply advanced analytics to augment existing data sources and design solutions for managing the budget.
	Recommend enhancements to existing systems to business needs by creating standard reports.	Demonstrate an understanding of data modeling and analytics concepts (queries, reporting, association of data sources).	Manage analytical process to deliver timely, insightful, and actionable analyses.

As Table 4.6 demonstrates, job requirements for the five professions included in our study overlap in a complex way. Our study sets out to sort out some parts of that complexity. In the next section, we present our methodology.

4.3.4 Methodology

4.3.4.1 Data Collection

Data were collected from the online source http://jobs.monster.com at three points in time: August 2015, July 2016, and July 2017. We searched for Job Opportunities focusing on the five Big Data professions (SA, DS, BDA, DA, and BA) described in the previous sections. The geographical location was set to retrieve job openings anywhere in the USA. The hiring companies/organizations belonged to a variety of industries. Each job description was pulled from

the website, and a document library was created. Three hundred job descriptions were collected each year, for a total of 900 job descriptions.

4.3.4.2 Text Analytics

The collected job qualification records were analyzed using Latent semantic analysis (LSA), which is a text analytic method that identifies text usage patterns to simulate word meaning (Deerwester et al. 1990; Landauer & Dumais, 1997). The analysis followed the guidelines in Evangelopoulos et al. (2012) and the steps in Kulkarni et al. (2014), which are listed below.

In step 1, we compiled a term-by-document frequency matrix, also known as the vector space model (Salton 1975). This is a data structure that quantifies unstructured text by recording the frequency of each term in each document. In order to finalize the set of documents and account for job description content effectively, job descriptions were split into individual passages that correspond to paragraphs, sections, or list items, likely to cover about one topic each. The resulting dataset consists of 8,986 such individual passages of unstructured text, which represent the columns in the term frequency matrix. The average passage size was 86 characters. In order to finalize the set of terms, we excluded the terms that appeared only once in the entire collection, as well as trivial English words (stopwords), such as and, the, therefore, etc. The spelling of terms that appear in different spelling styles was standardized, and acronyms were spelled out (e.g., BS to Bachelor's degree). The final list of terms included terms that appeared in at least four passages in the entire data set. These terms were conflated with term stemming, to merge terms that share a common stem, for a total of 2,519 unique stemmed terms. The raw term frequencies were transformed by applying an inverse document frequency (TF-IDF) term frequency transformation function. This weighting method discounts the occurrence of high-frequency terms and promotes

the occurrence of low-frequency terms, making it easier to identify abstract concepts (Evangelopoulos et al. 2012).

In step 2, the transformed term frequency matrix A was subjected to singular value decomposition, or $A = U\Sigma V^T$, where U are the term eigenvectors, V is the passage eigenvectors, and Σ is a diagonal matrix of singular values (i.e., square roots of common eigenvalues between terms and passages).

In step 3, a scree plot was used to examine the eigenvalues, in order to determine how many topics should be extracted. The scree plot is shown in Figure 4.3. One obvious choice would be to extract three very broad topics. However, since the intention was to obtain some additional information, we opted for the next "elbow point" of the scree plot, at nine topics. The log-likelihood ratio test for dimensionality detection (Zhu and Ghodsi, 2006), performed on eigenvalues 4 to 23, verified k = 9 as the dimensionality estimate (observed value of the test statistic $Q_n = 37.68$, p-value = 0.0022).

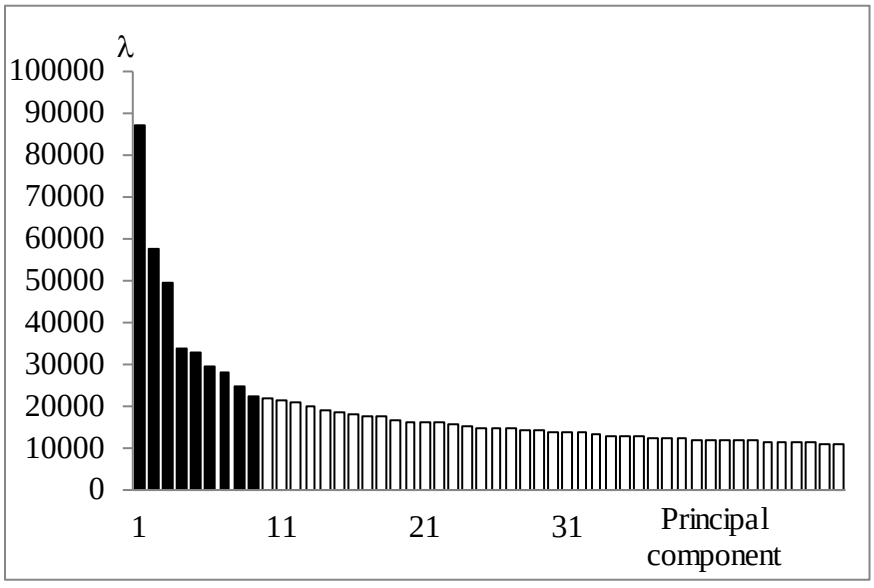

Figure 4.3: Scree plot of eigenvalues

In step 4, we labeled the topics using high-loading terms and high-loading documents. Topic labels and related high-loading terms and documents are shown in Table 4.7. Two of the authors compared their own topic labels and reached consensus quickly and without controversy.

4.3.5 Results

Table 4.7 shows the nine-broad latent semantic topics that characterize Big Data professions qualifications: business experience (t1), communication skills (t2), professional credentials (t3), Big Data skills (t4), professional skills (t5), programming skills (t6), statistical skills (t7), problem solving skills (t8), and database skills (T9).

Table 4.7: Topic extraction

Topic	Topic Label	High-loading terms	#Docs
T1	Business Experience	+year,+experience,+experience,minimum,+business	1140
T2	Communication Skills	+communication,+skill,written,excellent,verbal	654
T3	Professional Credentials	+degree,+field,science,related,computer	772
T4	Big Data Skills	data,Big Data,big,Hadoop,+experience	1266
T5	Professional Skills	+ability,+work,+demonstrate,+team,+project	1200
T6	Database Skills	sql,knowledge,+database,relational,+sas institute	1035
T7	Statistics & Mach. Learn. Skills	+analysis,statistical,knowledge,+business,learning, machine	954
T8	Problem Solving Skills	strong,+solve,+problem,+skill,analytical	823
T9	Programming skills	python,+language,+experience,java,+sas institute	772

4.3.5.1 Correspondence Analysis

In order to visualize the contingency table of qualifications by job categories, we conducted correspondence analysis, which projected rows and columns of the contingency table on a new space of principal components (Lee, 1996) and produced a correspondence map. Correspondence maps are visualization tolls that have been heavily used in marketing to identify associations between brands and their attributes (Higgs, 1990; Whitlark & Smith, 2001). The correspondence

map in Figure 4.4 visualizes the projection of the row (qualifications) and the column (jobs) on the first two components. With the research dimensions and the jobs presented on the same plot, we were able to see how each job has dynamically positioned itself in the qualification landscape and show the dynamic movements of the five jobs in the past three years.

Tables 4.8 and 4.9 show the coordinates of the row and column categories (jobs and topics, respectively) in the space defined by the first two principal components. The first two principal component dimensions explain 78% of variance in the contingency table. Including the third component, the explained variance increases to 90%. Therefore, we consider the first three components adequate for the purpose of representing the row and column categories. Breaking down the overall proportion of explained variance, 0.90, by individual variables, as shown in Tables 4.8 and 4.9 (quality column), the proportions of variance explained by the three components are 0.80 or higher for most job categories (rows) and qualification topics (columns), with many of them in the 0.90s.

Component 1 appears to be soft skills versus computational skills dimension. Referring to the jobs, Component 1 has BA and DA on one end and BD & DS on the other end. This confirms the "soft skills vs. computational skills" nature of component 1. Component 2 has statistics on one side and Big Data skills on the other side. Therefore, component 2 appears to be a statistical vs. computational skills dimension. Component 3 has programming and problem-solving skills on one side and database skills and Big Data skills on the other side. Therefore, component 3 appears to be a skills vs. experience dimension.

A careful observation of Tables 4.8 and 4.9, as well as Figure 4.4, helped us track the dynamics of the five professions. BA remains extremely focused on soft skills across Component 1, slightly switches its focus from computational skills to statistics across Component 2, and moves

considerably from experience to skills across Component 3. BD remains extremely focused on computational skills across both Components 1 and 2 and moves considerably from experience to skills across Component 3. DA remains relatively focused on soft skills across Component 1, remains focused on computational skills across Component 2, and moves considerably from experience to skills across Component 3. DS remains extremely focused on computational skills across Component 1, remains relatively focused on statistical skills across Component 2, and makes a relatively small move towards experience across Component 3. SA remains somewhere in the middle between soft skills and computational skills across Component 1, remains relatively focused on statistical skills across Component 2, and has moved from experience to skills across Component 3.

Table 4.8: The contingency table

T01	T02	T03	T04	T05	T06	T07	T08	T09	Job	Qualification Topic
157	59	67	72	166	77	45	100	39	BA2015	Business Experience
59	20	29	57	31	47	28	18	31	BD2015	Communication Skills
92	38	52	97	71	76	50	58	28	DA2015	Professional Credentials
98	43	80	135	87	98	115	84	131	DS2015	Big Data Skills
42	25	41	35	38	35	47	24	45	SA2015	Professional Skills
76	58	50	28	111	42	39	74	6	BA2016	Database Skills
89	28	56	140	54	81	43	44	81	BD2016	Statistics & Mach. Learn. Skills
80	49	49	98	112	86	45	52	24	DA2016	Problem Solving Skills
67	29	57	104	57	51	85	39	74	DS2016	Programming skills
55	43	52	48	90	58	107	49	46	SA2016	
70	60	36	30	111	73	47	54	22	BA2017	
54	50	40	144	64	76	64	64	58	BD2017	
58	66	43	93	89	110	58	66	21	DA2017	
80	34	59	111	60	46	91	31	104	DS2017	
63	52	61	74	59	79	90	66	62	SA2017	

BA, business analytics professionals; BD, Big Data analytics professionals; DA, data analysts; DS, data scientists; SA, statistical analysts.

Table 4.9: Job categories on the principal component space

			Component Coordinates		
ID	Name	Quality	Comp. 1	Comp. 2	Comp. 3
1	BA2015	0.985	-0.325	0.017	0.21
2	BD2015	0.603	0.106	0.151	0.091
3	DA2015	0.667	-0.065	0.145	0.021
4	DS2015	0.865	0.234	-0.071	0.022
5	SA2015	0.804	0.135	-0.188	0.043
6	BA2016	0.939	-0.483	-0.104	0.051
7	BD2016	0.982	0.25	0.214	0.089
8	DA2016	0.819	-0.177	0.143	-0.029
9	DS2016	0.925	0.263	-0.062	0.022
10	SA2016	0.894	-0.025	-0.283	-0.108
11	BA2017	0.896	-0.368	-0.081	-0.046
12	BD2017	0.796	0.145	0.166	-0.136
13	DA2017	0.979	-0.168	0.126	-0.229
14	DS2017	0.944	0.346	-0.1	0.096
15	SA2017	0.731	0.074	-0.116	-0.112

Table 4.10: Qualification topics on the principal component space

			Component Coordinates		
ID	Name	Quality	Comp. 1	Comp. 2	Comp. 3
1	Business Experience	0.92	-0.107	0.05	0.203
2	Communication Skills	0.879	-0.245	-0.032	-0.139
3	Professional Credentials	0.525	0.037	-0.08	0.057
4	Big Data Skills	0.957	0.258	0.236	-0.035
5	Professional Skills	0.926	-0.332	-0.051	0.047
6	Database Skills	0.721	-0.052	0.125	-0.137
7	Statistics & Mach. Learn. Skills	0.95	0.179	-0.254	-0.122
8	Problem Solving Skills	0.633	-0.209	-0.009	-0.024
9	Programming skills	0.962	0.493	-0.12	0.106

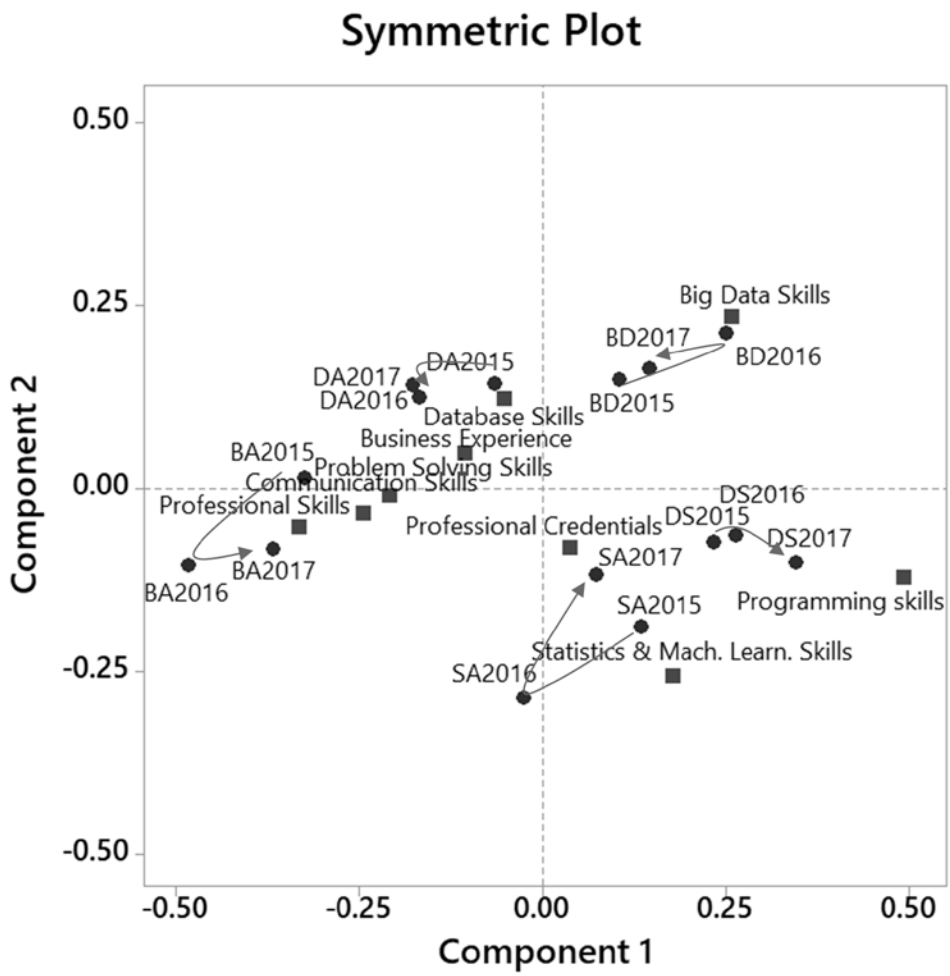

Figure 4.4: Correspondence map showing a two-dimensional projection of the 15-by-9 matrix (BA, business analytics professionals; BD, Big Data analytics professionals; DA, data analysts; DS, data scientists; SA, statistical analysts)

4.3.6 Discussion

Our field study of 900 job announcements with nearly 9000 paragraphs of text and our text analytic treatments confirm what the experts have already stated: that in the current environment, whether we are referring to statistical analysts (SA), Big Data analytics professionals (BDA), data scientists (DS), data analysts (DA), or business analytics professionals (BA), analytics, IT, communication, and business skills are all important abilities of the competitive professional (De Veaux et al. 2017).

Comparing job descriptions in 2015 to job descriptions in 2016 and 2017, our results point to a relatively stable positioning of the five professions in the space of skills and qualifications. Overall, all five professions included in our study remain close to their respective core skills. BA remains close to communication and professional skills, DA remains close to database skills, BD remains close to Big Data skills, SA remains close to statistics and machine learning skills, and DS remains at a position that balances statistics and programming skills. Interestingly, based on the insights gained from our analysis, BA, BD, DA, and SA have all shifted their focus from experience to skills over the last three years. This may indicate that, as of 2017, as these positions were in high demand, and the job market was willing to settle for skills, in the absence of job seekers with more experience, since those professional that had the experience were already employed. Specifically for SA, this trend is highlighted by the highly-ranked listing of the profession of statistician by U.S. News (2017). Regarding DS, where the opposite trend towards experience was observed, the explanation might be that data science, being the youngest of the five professions included in our study, is slowly getting to the point where the job market can finally find some professionals with experience; therefore, jobs start asking for it. Further, a slight move of the DS job category away from the professional credentials requirement may indicate some frustration on the part of employers who cannot find candidates with those credentials and start asking more specifically for statistical and programming skills.

Our results confirm the statement by Hoerl et al. (2010) and Hardin et al. (2015) that there is a need for communication skills and statistical thinking in the context of working with data. They also confirm the spirit of the entire data science curriculum development initiative in De Veaux et al. (2017), where data science was approached as a balanced mix of statistics and computer science. However, our results only partially confirm Sharda et al. (2018, p. 460), where

a framework of skills that define a data scientist includes communication and interpersonal skills, curiosity and creativity, IT and social media skills, programming skills, data management skills, and domain expertise. We see a need to include statistical skills in that list. In fact, on the very next page of the same reference, Sharda et al. (2018, p. 461) display an insert with a typical job post for data scientists, where “appropriate statistical techniques” are clearly listed.

4.3.7 Conclusion

In summary, we continue the American Statistician’s long tradition of articles and recommendations about the statistics profession (e.g., Fox, 2010; ASA, 2014). Our paper builds upon a long tradition of related studies and commentaries. Rather than relying on subjective expertise, our study has examined the job market to distill a space of skill and qualification dimensions from job announcements related to five Big Data professions. We have placed the five professions on a map created by these dimensions and have observed their recent trends. Overall, requirements for statisticians and data scientists remain far away from soft skills, such as professional, communication, and business problem-solving skills. The requirements for data scientists have settled at a balance of statistical, machine learning, and programming skills, confirming the curriculum development guidelines in De Veaux et al. (2017).

CHAPTER 5

CONCLUSION

Big Data has become a major focus in many sectors and across numerous applications within sectors. The premises and utilization of methods have a significant impact on the resulting decision-making. However, the lack of an established definition and common terms as applied to Big Data has caused a lack of consistency in both academic research and applied applications. In order to clarify the definition and common terms applied to Big Data, this research examined the definition and the profession and explored the influence of Big Data in one specific application to provide insight into this emerging research arena.

This research recommends a new way to look at Big Data and one source of such data, social media. This recommendation is well-suited to the theoretical and methodological foundations of Big Data and addresses an increasing demand for more powerful Big Data analysis from an academic prospective. Essay 1 provided a strategic overview of the untapped potential of big social media data, explained its challenges and opportunities for aspiring organizations, and made recommendations on how companies can exploit social media data analysis to make better decisions—decisions that embrace the relevant social qualities of their customers and their related ecosystem.

Essay 2 provided a better understanding of the influence of social media during the 2016 American presidential election and developed an SM influence model to examine individuals' attitudes toward participating in social media (SM) discussions that might influence their decision in choosing between the two presidential election candidates, Donald Trump and Hilary Clinton, by integrating social influence theory, TAM2, and media richness theory. The research provided insights about the influence of SM on voting decisions and indicated that perceived social

influence is a major factor because it significantly influences SM community identification, while SM community identification strongly influences individuals' attitudes. This research indicated that millennials are influenced by people whom they follow on social media sites, and what they share and discuss has shaped millennials' vote decisions.

Essay 3 helped to clarify the skills required for Big Data professionals for the joint benefit of the job market where they will be employed and academia, where such professionals will be prepared in data science programs, suggesting improvements in the entire process of preparing and recruiting for Big Data positions. The results also addressed the need for clarification of such overlapping subjects not only in industry but also in the academic environment. The work examined the job market from 2015-2017 to distill a space of skill and qualification dimensions from job announcements related to five Big Data professions. This study placed the five professions on a map created by these dimensions, observed their recent trends, and showed the differences and commonalities among job description elements for SA, BDA, DS, DA, and BA. It also illustrated the dynamic behavior of these differences and commonalities in recent years.

Future research will focus on using different methodologies to compare the results. The collected job qualification records will be analyzed using linear discriminant analysis (LDA), which is an example of dimensionality reduction, (e.g., principal components analysis, analysis of variance, and regression analysis). In this technique, the analyst looks for the areas in the data that have the greatest variance and then projects the data onto them (Welling, 2005).

Cumulatively, the three works provided in this dissertation yielded a new way to look at both Big Data and Big Data analysis. This research provides new theoretical and methodological support that refines the definition and methodologies associated with Big Data and provides insights for Big Data analysis as well as for academic research.

REFERENCES

Ackoff, R. L. (1989). From data to wisdom. *Journal of applied systems analysis*, 16(1), 3-9.

Actionable Tips to Analyze Unstructured Data, (). Available at https://www.michiganstateuniversityonline.com/resources/business-analytics/actionable-tips-to-analyze-unstructured-data/#.WX-GydPyuCQ

Aggarwal, C. C. (2011). An introduction to social network data analytics. *Social network data analytics*, 1-15.

American Statistical Association Undergraduate Guidelines Workgroup (ASA) (2014), 2014 Curriculum Guidelines for Undergraduate Programs in Statistical Science, Alexandria, VA: *American Statistical Association*. Available at http://www.amstat.org/education/curriculumguidelines.cfm.

Ancu, M., & Cozma, R. (2009). Myspace politics: uses and gratifications of befriending candidates. *Journal of Broadcasting & Electronic Media*, 53, 567–583.

Ariker, M., McGuire, T., & Perry, J. (2013). Five roles you need on your Big Data team. *Harvard Business Review*, July 22, 2013. Available at https://hbr.org/2013/07/five-roles-you-need-on-your-bi.

Balust, J., & Macario, A. (2009). Can anesthesia information management systems improve quality in the surgical suite? *Current Opinion in Anesthesiology*, 22(2), 215-222.

Barbaro, M., (2016). How Did the Media — How Did We — Get This Wrong? Available at https://www.nytimes.com/2016/11/09/podcasts/election-analysis-run-up.html

Barbier, G., & Liu, H. (2011). Data mining in social media. In C. C. Aggarwal (Ed.), *Social network data analytics* (pp. 327–352). United States: Springer.

Bates, M. J. (1999). The invisible substrate of information science. *Journal of the American Society for Information Science*, 50(12), 1043-1050.

Beeson, I., & Chelin, J. (2006). Information systems meets information science. *Innovation in Teaching and Learning in Information and Computer Sciences*, 5(2), 1-6.

Belkin, N. J., & Robertson, S. E. (1976). Information science and the phenomenon of information. *Journal of the American Society for Information Science*, 27(4), 197-204.

Bellinger, G., Castro, D., & Mills, A. (2004). *Data, information, knowledge, and wisdom*.

Bolton, R. N., Parasuraman, A., Hoefnagels, A., Migchels, N., Kabadayi, S., Gruber, T., & Solnet, D. (2013). Understanding generation Y and their use of social media: a review and research agenda. *Journal of Service Management*, 24(3), 245-267.

Borko, H. (1968). Information science: what is it? *American documentation*, 19(1), 3-5.

Boström, H., Andler, S. F., Brohede, M., Johansson, R., Karlsson, A., Van Laere, J., & Ziemke, T. (2007). On the definition of information fusion as a field of research.

Brown, M. S. (2016, May 31). Big data analytics and the next president: how microtargeting drives today's campaigns. Available at http://www.forbes.com/sites/metabrown/2016/05/29/big-data-analytics-and-the-next-president-how-microtargeting-drives-todays-campaigns/#72cf44301400

Buckland, M. (1999). The landscape of information science: The American Society for Information Science at 62. *Journal of the American Society for Information Science*, 50(11), 970-974.

Buckland, M. (2012). What kind of Science can Information Science be? *Journal of the American Society for Information Science and Technology*, 63(1), 1-7.

Burke, F., (2013). Social media vs. social networking. Available at https://www.huffingtonpost.com/fauzia-burke/social-media-vs-social-ne_b_4017305.html

Campos, Luiz Fernando de Barros. (2013). Social information. *Transinformação*, 25(2), 151-157. Available at https://dx.doi.org/10.1590/S0103-37862013000200006

Cha, M., Haddadi, H., Benevenuto, F., & Gummadi, P. K. (2010). Measuring user influence in twitter: The million follower fallacy. *ICWSM*, 10(10-17), 30.

Chamandy, N., Muralidharan, O., & Wager, S. (2015). Teaching statistics at google-scale. *The American Statistician*, 69(4), 283–291.

Chen, C. P., & Zhang, C. Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information Sciences*, 275, 314-347.

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business intelligence and analytics: From Big Data to big impact. *MIS quarterly*, 36(4), 1165-1188.

Chen, M., Ebert, D., Hagen, H., Laramee, R. S., Van Liere, R., Ma, K. L., & Silver, D. (2009). Data, information, and knowledge in visualization. *Computer Graphics and Applications, IEEE*, 29(1), 12-19.

Chen, M., Mao, S., & Liu, Y. (2014). Big Data: a survey. *Mobile Networks and Applications*, 19(2), 171-209.

Cheong, F., & Cheong, C. (2011). Social media data mining: a social network analysis of tweets during the 2010-2011 Australian floods. *PACIS*, 11, 46-46.

Cheung, C. M., Chiu, P. Y., & Lee, M. K. (2011). Online social networks: Why do students use Facebook? *Computers in Human Behavior*, 27(4), 1337-1343.

Chin, W. W. (1998). The partial least squares approach to structural equation modeling. *Modern methods for business research*, 295(2), 295-336.

Cleary, R., & Woolford, S. (2010). The business of desire and fear. *The American Statistician*, 64(1), 21–22.

Cleveland, W. S. (2001). Data science: an action plan for expanding the technical areas of the field of statistics. *International statistical review*, 69(1), 21–26.

Constine, J., (2017). Facebook now has 2 billion monthly user and responsibility. Available at https://techcrunch.com/2017/06/27/facebook-2-billion-users/

Costello, J., (2017). America runs on dunkin'. dunkin' runs on 13 million fans. Available at https://www.salesforce.com/customer-success-stories/dunkin-brands/

Cukier, K., & Mayer-Schoenberger, V. (2013). Rise of Big Data: how it's changing the way we think about the world, *The Foreign.*, 92, 28.

Davenport, T. (2014). Big Data at work: dispelling the myths, uncovering the opportunities. *Harvard Business Review Press*.

Davenport, T. H., & Dyché, J. (2013). Big Data in Big Companies. *International Institute for Analytics*, 1-31.

Davis, F. D. "Perceived usefulness, perceived ease of use, and user acceptance of information technology," *MIS Quarterly*, 13, 1989, pp. 319-340.

Dbmesser. (2011). IS-Relationships [Chart]. Available at https://commons.wikimedia.org/wiki/File:IS-Relationships-Chart.jpg

De Mauro, A., Greco, M., & Grimaldi, M. (2016). A formal definition of Big Data based on its essential features. *Library Review*, 65(3), 122-135.

De Veaux, R.D., Agarwal, M., Averett, M., Baumer, B.S., Bray, A., Bressoud, T.C., Bryant, L., Cheng, L.Z., Francis, A., Gould, R., Kim, A.Y., Kretchmar, M., Lu, Q., Moskol, A., Nolan, D., Pelayo, R., Raleigh, S., Sethi, R.J., Sondjaja, M., Tiruviluamala, N., Uhlig, P.Z., Washington, T.M., Wesley, C.L., White, D., Ye, P. (2017). Curriculum guidelines for undergraduate programs in data science. *Annual Review of Statistics and its Application*, 4, 2.1–2.16.

Dean, J., & Ghemawat, S. (2008). MapReduce: Simplified data processing on large clusters. Communications of the ACM, 51(1), 107-113.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. Journal of the American society for information science, 41(6), 391.

DeMers, J., (2014). The top 10 benefits of social media marketing. Available at https://www.forbes.com/sites/jaysondemers/2014/08/11/the-top-10-benefits-of-social-media-marketing/#2ca507bf1f80

Dennis, A. R., & Valacich, J. S. (1999, January). Rethinking media richness: Towards a theory of media synchronicity. In Systems Sciences, 1999. HICSS-32. Proceedings of the 32nd Annual Hawaii International Conference on (pp. 10-pp). IEEE.

Dholakia, U. M., Bagozzi, R. P. and Pearo, L. K. (2004). A social influence model of consumer participation in network- and small-group-based virtual communities. International Journal of Research in Marketing, 21: 241-263.

Digital display advertising revenue by company (2017). Available at http://www.journalism.org/chart/digital-display-advertising-revenue-by-company/

Dimitrova, D. V., Shehata, A., Strömbäck, J., & Nord, L. W. (2014). The effects of digital media on political knowledge and participation in election campaigns: Evidence from panel data. Communication Research, 41(1), 95-118.

Edmunds, A., & Morris, A. (2000). The problem of information overload in business organisations: a review of the literature. International Journal of Information Management, 20(1), 17-28.

Effing, R., Van Hillegersberg, J., & Huibers, T. (2011). Social media and political participation: are Facebook, Twitter and YouTube democratizing our political systems?. In International Conference on Electronic Participation (pp. 25-35). Springer Berlin Heidelberg.

Ellis, D., Allen, D., & Wilson, T. (1999). Information science and information systems: Conjunct subjects disjunct disciplines. Journal of the Association for Information Science and Technology, 50(12), 1095.

Ellison, N. B. (2007). Social network sites: Definition, history, and scholarship. Journal of Computer-Mediated Communication, 13(1), 210–230.

Erlandson, R. (2013). Social Media and Social Networking. *Technology for Small and One-Person Libraries: A LITA Guide*, 21, 85.

Evangelopoulos, N., Zhang, X. & Prybutok, V. (2012) Latent semantic analysis: five methodological recommendations. European Journal of Information Systems, 21, 70–86.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. AI magazine, 17(3), 37.

Feldman, R., & Sanger, J. (2007). The text mining handbook: advanced approaches in analyzing unstructured data. Cambridge university press.

Finlay, P. N., & Forghani, M. (1998). A classification of success factors for decision support systems. The Journal of Strategic Information Systems, 7(1), 53-70.

Fleck, J., Johnson-Migalski, L., (2015). The impact of social media on personal and professional lives: an Adlerian perspective. J. Individ. *Psychol*. 71(2), 135–142

Ford, F. N. (1985). Decision support systems and expert systems: a comparison. Information & Management, 8(1), 21-26.

Fox, D. R. (2010). Desired and Feared—Quo vadis or Quid agis? The American Statistician, 64(1), 6–9.

French, C. (1996). Data Processing and Information Technology. *Cengage Learning EMEA*.

Friedman, J. H. (2001). The role of statistics in the data revolution? *International Statistical Review*, 69(1), 5–10.

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big Data concepts, methods, and analytics. *International Journal of Information Management*,35(2), 137-144.

Gartner (2017). Business Analytics. Gartner IT Glossary. Available at http://www.gartner.com/it-glossary/business-analytics/.

Gil de Zu´n˜ iga, H., Puig, E., & Rojas, H. (2009). Weblogs traditional sources online & political participation: An assessment of how the Internet is changing the political environment. *New Media & Society*, 11, 553–574.

Goodrich, K., & De Mooij, M. (2014). How 'social'are social media? A cross-cultural comparison of online and offline purchase decision influences. *Journal of Marketing Communications*, 20(1-2), 103-116.

Groshek, J., & Dimitrova, D. (2013). A cross-section of political involvement, partisanship and online media in Middle America during the 2008 presidential campaign. *Atlantic Journal of Communication*, 21, 108–124.

Gruhl, D., Guha, R., Kumar, R., Novak, J., & Tomkins, A. (2005). The predictive power of online chatter. In Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining (pp. 78-87). *ACM*.

Gruzd, A., & Wellman, B. (2014). Networked influence in social media: introduction to the special issue. *American Behavioral Scientist*, 58(10) 1251–1259.

Gruzd, A., Jacobson, J., Mai, P., Ruppert, E., & Murthy, D. (2016). Introduction to the 2016 International Conference on Social Media and Society. *In Proceedings of the 7th 2016 International Conference on Social Media & Society* (p. 1). ACM.

Gundecha, P., & Liu, H. (2012). Mining social media: A brief introduction. *Tutorials in Operations Research*, 1(4)

Hadoop, A. (2009). Hadoop. 2009-03-06. Available at http://hadoop. apache. org.

Hair Jr, J. F., Hult, G. T. M., Ringle, C., & Sarstedt, M. (2016). *A primer on partial least squares structural equation modeling (PLS-SEM).* Sage Publications.

Hair, J. F. J., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate data analysis* (7th ed.). Prentice Hall, Upper Saddle River, NJ.

Hall, S. (2012). Statistical data analyst job description. Available at http://work.chron.com/statistical-data-analyst-job-description-15901.html.

Hand, D. J. (2007). Principles of data mining. *Drug safety*, 30(7), 621-622

Hardin, J., Hoerl, R., Horton, N. J., Nolan, D., Baumer, B., Hall-Holt, O., & Ward, M. D. (2015). Data science in statistics curricula: Preparing students to "Think with Data". *The American Statistician*, 69(4), 343–353.

He, W., Wu, H., Yan, G., Akula, V., & Shen, J. (2015). A novel social media competitive analytics framework with sentiment benchmarks. *Information & Management*, 52(7), 801-812.

He, W., Zha, S., & Li, L. (2013). Social media competitive analysis and text mining: A case study in the pizza industry. *International Journal of Information Management*, 33(3), 464-472.

Hemley, D., (2013). 26 Tips to Create a Strong Social Media Content Strategy. Available at http://www.socialmediaexaminer.com/26-tips-to-create-a-strong-social-media-content-strategy/

Henseler, J., Ringle, C. M., & Sinkovics, R. R. (2009). The use of partial least squares path modeling in international marketing. *In New challenges to international marketing* (pp. 277-319). Emerald Group Publishing Limited.

Higgs, N.T. (1990). Practical and innovative uses of correspondence analysis. *Journal of the Royal Statistical Society Series D (The Statistician)*, 40, Special Issue: Survey Design, Methodology and Analysis, 183–194.

Hinchcliffe, D., (2012). Ten examples of extracting value from social media using Big Data. Available at http://www.zdnet.com/pictures/ten-examples-of-extracting-value-from-social-media-using-big-data/7/

Hoerl, R. W., & Snee, R. D. (2010). Moving the statistics profession forward to the next level. *The American Statistician*, 64(1), 10–14.

Holmes, R. (2015). 5 trends that will change how companies use social media in 2016. Available at https://www.fastcompany.com/3054347/5-trends-that-will-change-how-companies-use-social-media-in-2016

IBM, (2013). The Evolution of Big Data [Infographic]. Available at http://www.ibmbigdatahub.com/infographic/evolution-big-data

Inmon, W. H., Imhoff, C., & Battas, G. (1999). *Building the operational data store* (Vol. 8). John Wiley.

Issid, J., (2017). The evolution of job titles. Available at https://www.monster.ca/career-advice/article/new-job-titles-in-the-market-ca

Jamie, (2017). Available at https://makeawebsitehub.com/social-media-sites/

Jasonkarpf, J. (2016). Clinton and Pollsters Lose. Available at https://funkyadjunct.com/2016/11/12/clinton-and-pollsters-lose/

Johnstone, I., & Roberts, F. (2014). Data Science at NSF. NSF, 7(29), 2014. Available at https://www.nsf.gov/attachments/130849/public/Stodden-StatsNSF.pdf.

Kallus, N. (2014). Predicting crowd behavior with big public data. *In Proceedings of the 23rd International Conference on World Wide Web* (pp. 625-630). ACM.

Kandel, S., Paepcke, A., Hellerstein, J. M., & Heer, J. (2012). Enterprise data analysis and visualization: An interview study. *Visualization and Computer Graphics, IEEE Transactions on*, 18(12), 2917-2926.

Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business horizons*, 53(1), 59-68.

Katal, A., Wazid, M., & Goudar, R. H. (2013). Big Data: issues, challenges, tools and *good practices. In* Contemporary Computing (IC3), 2013 Sixth International Conference on (pp. 404-409). IEEE.

Keen, P. G. (1987). Decision support systems: the next decade. *Decision Support Systems*, 3(3), 253-265.

Keeter, S., Horowitz, J., Tyson A. (2008). Young Voters in the 2008 Election. Available at http://www.pewresearch.org/2008/11/13/young-voters-in-the-2008-election/

Keim, D., Qu, H., & Ma, K. L. (2013). Big-data visualization. Computer Graphics and Applications, *IEEE*, 33(4), 20-21.

Kelman, H. C. (1974). Further thoughts on the processes of compliance, identification, and internalization. *Perspectives on Social Power*. J. T. Tedeschi. Chicago, Aldine Press: 126-171.

Kim, G. H., Trimi, S., & Chung, J. H. (2014). Big-data applications in the government sector. Communications of the ACM, 57(3), 78-85.

Kimball, R., & Ross, M. (2011). The data warehouse toolkit: the complete guide to dimensional modeling. John Wiley & Sons.

Klein, M., Methlie, L.B., 1990. *Expert Systems. A Decision Support Approach with Applications in Management and Finance*. Addison Wesley Publishing Company.

Kohut, A., Keeter, S., Doherty, C., & Dimock, M. (2008). Social networking and online videos take off: Internet's broader role in campaign 2008. *TPR Center, The PEW research center*.

Kulkarni, S., Apte, U. & Evangelopoulos, N. (2014) The use of latent semantic analysis in operations management research. *Decision Sciences*, 45, 971–994.

Kushin, M. J., & Yamamoto, M. (2010). Did social media really matter? College students' use of online media and political decision making in the 2008 election. *Mass Communication and Society*, 13(5), 608-630.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2), 211.

LaValle, S., Lesser, E., Shockley, R., Hopkins, M. S., & Kruschwitz, N. (2013). Big Data, analytics and the path from insights to value. *MIT Sloan Management Review* 52(2), 21-32.

Leary, M. R., & Kowalski, R. M. (1990). Impression management: A literature review and two-component model. *Psychological bulletin*, 107(1), 34.

Lee, B. L. (1996). *Correspondence analysis*. LL Thurstone Psychometric Laboratory Research Memorandum.

Levine, S., (2011). 8 things you can do to evolve your career. Available at http://www.businessinsider.com/8-tips-company-changes-james-marshall-reilly-2011-11

Life Project. Washington, DC: Pew Trust. Available at http://www.pewinternet.org/2008/06/15/the-internet-and-the-2008-election/

Liu, S. H., Liao, H. L., & Pratt, J. A. (2009). Impact of media richness and flow on e-learning technology acceptance. *Computers & Education*, 52(3), 599-607.

Manovich, L. (2011). Trending: The promises and the challenges of big social data. *Debates in the digital humanities*, 2, 460-475.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). Big Data: The next frontier for innovation, competition (p. 9). and productivity. *McKinsey Global Institute*. Available at http://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation

March, S. T., & Smith, G. F. (1995). Design and natural science research on information technology. *Decision Support Systems*, 15(4), 251-266.

Markman, J. (2016). Big Data and the 2016 election. Available at http://www.forbes.com/sites/jonmarkman/2016/08/08/big-data-and-the-2016-election/#555a08cd46d7

Mason, R. L. (2004). Does the statistics profession have an identity crisis? *Journal of the American Statistical Association*, 99(465), 1-6.

Mayer-Schönberger, V., & Cukier, K. (2013). *Big Data: A revolution that will transform how we live, work, and think*. New York: Houghton Mifflin Harcourt.

McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D. J., & Barton, D. (2012). Big Data. The management revolution. *Harvard Bus Review*, 90(10), 61-67.

McHugh, M. L. (2013). The chi-square test of independence. *Biochemia medica*, 23(2), 143-149.

McKinney Jr, E. H., & Yoos, C. J. (2010). Information about information: A taxonomy of views. *MIS quarterly*, 329-344.

Mell, P., & Grance, T. (2011). The NIST definition of cloud computing. NIST Special Publication 800-145

Mello, R., Leite, L. R., & Martins, R. A. (2014). Is Big Data the next big thing in performance measurement systems? *In IIE Annual Conference. Proceedings* (p. 1837). Institute of Industrial Engineers-Publisher.

Michaels, V. (2017). Did social media influence the U.S. election? Available at https://www.engadget.com/2017/01/02/did-social-media-influence-the-u-s-election/

Monarch, I. (2000). Information science and information systems: Converging or diverging. *In Proceedings of the 28th Annual Conference of the Canadian Association for Information.*

Moy, P., Manosevitch, E., Stamm, K. R., & Dunsmore, K. (2005). Linking dimensions of Internet use and civic engagement. *Journalism & Mass Communication Quarterly*, 82, 571-586.

MSNBC. (2017). Tweeter-in-chief: how twitter is a factor in Donald Trump's presidency the 11th hour MSNBC. Available at https://cdn.eblnews.com/video/tweeter-chief-how-twitter-factor-donald-trumps-presidency-11th-hour-msnbc-90305

Nagle, S. (2016). Available at https://www.rw3.com/understanding-the-difference-scorecards-vs-dashboards/

Nicholls, D. F. (2001). Future directions for the teaching and learning of statistics at the tertiary level. *International Statistical Review/Revue Internationale de Statistique*, 11–15.

NSF (Natl. Sci. Found.). 2014. Data Science at NSF Draft Report of StatSNSF Committee: Revisions Since January MPSAC Meeting. April. Available at https://www.nsf.gov/attachments/130849/public/Stodden-StatsNSF.pdf

Number of global social network users 2010-2021, (2017). Available at https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/

Nunnally, J. C. (1978). *Psychometric theory*. New York, NY: McGraw-Hill.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory*. New York: McGraw.

Olavsrud, T. (2014). C-Level Executives Seeing Big Results from Big Data

Pew (2017). Available at http://www.pewinternet.org/fact-sheet/social-media/

Piatetsky, G. (2016). Trump, Failure of Prediction, and Lessons for Data Scientists, KDnuggets. Available at http://www.kdnuggets.com/2016/11/trump-shows-limits-prediction.html

Power, D. J. (2002). Decision support systems: concepts and resources for managers. *Studies in Informatics and Control*, 11(4), 349-350.

Power, D. J., & Phillips-Wren, G. (2011). Impact of social media and Web 2.0 on decision-making. *Journal of decision systems*, 20(3), 249-261.

Prensky, M. (2001). "Digital natives, digital immigrants", *On the Horizon*, Vol. 9 No. 5

Prewitt, T., (2017). The benefits of social media: how to choose strategically. Available at https://www.envision-creative.com/benefits-of-social-media-choosing-strategically/

Rainier, K. R., & Cegielski, C. G. (2012). *Introduction to Information Systems: Supporting and Transforming Business*. Hoboken.

Rajaraman, A., and Ullman, J., (2011). *Mining of Massive Datasets*, Cambridge University Press, 2011.

Rashotte, L. (2007). Social influence. *The blackwell encyclopedia of social psychology*, 9, 562-563.

Rasp, S., (2016). The value in unstructured data. BDW. Available at http://blog.bigdataweek.com/2016/08/01/value-unstrucutred-data/

Real-time results for the presidential race. (2016). Available at https://www.washingtonpost.com/2016-election-results/us-presidential race/?utm_term=.6be1e5fd9d30

Rijmenam, M. (2016). Big Data Analyst Profile. Available at https://datafloq.com/read/job-description-big-data-analyst/198

Ringle, C. M., Wende, S., & Will, A. (2005). *SmartPLS 2.0(beta).* Hamburg, Germany.

Rouse, M., (2016). Available at http://searchcloudcomputing.techtarget.com/definition/Software-as-a-Service

Russom, P. (2011). Big Data analytics. *TDWI Best Practices Report*, Fourth Quarter.

Russom, P. (2013). Managing Big Data. *TDWI Best Practices Report*, TDWI Research, 1–40.

Sago, B. (2010). The Influence of Social Media Message Sources on Millennial Generation Consumers. *International Journal of Integrated Marketing Communications*, 2(2).

Scarf, M., (2012). Social media and the Big Data explosion. Available at https://www.forbes.com/sites/onmarketing/2012/06/28/social-media-and-the-big-data-explosion/#72541a436a61

Schauer, P., (2015). 5 Biggest differences between social media and social networking. Available at https://www.socialmediatoday.com/social-business/peteschauer/2015-06-28/5-biggest-differences-between-social-media-and-social

Schmitz, J., & Fulk, J. (1991). Organizational colleagues, media richness, and electronic mail: A test of the social influence model of technology use. *Communication research*, 18(4), 487-523.

Schoen, H., Gayo-Avello, D., Takis Metaxas, P., Mustafaraj, E., Strohmaier, M., & Gloor, P. (2013). The power of prediction with social media. *Internet Research*, 23(5), 528-543.

Shah, C. (2017). Social Media and Social Networking. In Social Information Seeking (pp. 29-42). *Springer*, Cham.

Shah, D. V., Kwak, N., & Holbert, R. L. (2001). ‘‘Connecting’’ and ‘‘disconnecting’’ with civic life: Patterns of Internet use and the production of social capital. *Political Communication*, 18, 141-162.

Shannon, C.E., & Weaver, W. (1949). The mathematical theory of communication. *Urbana, IL*: University of Illinois Press.

Shannon, H., (2015). How to build a social media marketing strategy for your business. Available at http://www.addthis.com/blog/2015/09/11/how-to-build-a-social-media-marketing-strategy-for-your-business/#.WXdb_tPyuCQ

Sharda, R., Delen, D., and Turban, E. (2018), *Business Intelligence, Analytics, and Data Science: A Managerial Perspective*. Upper Saddle River, NJ: Pearson Education, 4th Edition.

Shigemi, N., Yamamoto, H., Tazaki, G., Yoshioka, M., & Kokubun, M. (2001). Structured data management system and computer-readable recording medium storing structured data management program. *U.S. Patent*, No. 6,314,434. Washington, DC: U.S. Patent and Trademark Office.

Siegel, E. (2013). *Predictive analytics: The power to predict who will click, buy, lie, or die*. John Wiley & Sons

Smith, A., & Rainie, L. (2008). The Internet and the 2008 election. Pew Internet & American

Smith, K., (2017). Marketing: 105 amazing social media statistics and facts. Available at https://www.brandwatch.com/blog/96-amazing-social-media-statistics-and-facts-for-2016/

Smith, M. A., Shneiderman, B., Milic-Frayling, N., Mendes Rodrigues, E., Barash, V., Dunne, C., & Gleave, E. (2009). Analyzing (social media) networks with NodeXL. *In Proceedings of the fourth international conference on Communities and technologies* (pp. 255-264). ACM.

Social network advertising revenue from 2014 to 2017 (in billion U.S. dollars) (2017). Available at https://www.statista.com/statistics/271406/advertising-revenue-of-social-networks-worldwide/

Sprague Jr, R. H. (1980). A framework for the development of decision support systems. *MIS Quarterly*, 1-26.

Stadd, A. (2014). Data analysts: what you'll make and where you'll make it | udacity. Available at http://blog.udacity.com/2014/11/data-analysts-what-youll-make.html.

Starr, R. (2013). Accenture study shows businesses flag Big Data as important for digital transformation. Available at http://www.big4.com/big4-thought-leader-interviews/accenture-study-shows-businesses-flag-big-data-important-digital-transformation/.

Stephens, M., Yoo, J., Mourão, R. R., Gutierrez, F. M., Baresch, B., & Johnson, T. J. (2016). The life of the tea party: differences between tea party and republican media use and political variables. *Atlantic Journal of Communication*, 24(3), 157-171.

Stieglitz, S., & Dang-Xuan, L. (2013). Social media and political communication: a social media analytics framework. *Social Network Analysis and Mining*, 3(4), 1277-1291.

Tang, J., & Liu, H. (2012). Unsupervised feature selection for linked social media data. *In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 904-912). ACM.

Troester, M. (2012). Big Data meets Big Data analytics white paper https://www.sas.com/content/dam/SAS/en_us/doc/whitepaper1/big-data-meets-big-data-analytics-105777.pdf

U.S. News (2017). Statistician Overview. Available at http://money.usnews.com/careers/best-jobs/statistician.

Using Data to Drive Social Media Decision-Making, (2016). Available at http://fleishmanhillard.com/2016/04/social-innovation/using-data-to-drive-social-media-decision-making/

Utz, S. (2009). The (potential) benefits of campaigning via social network sites. *Journal of Computer-Mediated Communication*, 14, 221–243.

Valenzuela, S. (2013). Unpacking the use of social media for protest behavior: The roles of information, opinion expression, and activism. *American Behavioral Scientist*, 57(7), 920-942.

Valenzuela, S., Park, N., & Kee, K. F. (2009). Is there social capital in a social network site? Facebook use and college students' life satisfaction, trust, and participation. *Journal of Computer-Mediated Communication*, 14, 875–901

Varon, E. (2012). Rethink your org chart for Big Data analytics teams. Available at http://data-informed.com/rethink-your-org-chart-for-big-data-analytics-teams/

Venkatesh, V., & Davis, F. D. (2000). A theoretical extension of the technology acceptance model: Four longitudinal field studies. *Management science*, 46(2), 186-204.

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly*, 425-478.

Venkatesh, V., Thong, J. Y. L., & Xu, X. (2012). Consumer acceptance and use of information technology: Extending the unified theory of acceptance and use of technology. *MIS Quarterly*, 36(1), 157–178

Walker, R., (2016). From Big Data to big profits. *Harvard Business Review*. Available at https://hbr.org/webinar/2016/03/from-big-data-to-big-profits

Wang, S., & Wang, H. (2010). Towards innovative design research in information systems. *Journal of Computer Information Systems*, 51(1), 11-18.

Webber, S. (2003). Information science in 2003: A critique. *Journal of Information Science*, 29(4), 311-330.

Welch, B., (2017). Data analyst. Available at https://www.prospects.ac.uk/job-profiles/data-analyst

Welling, M. (2005). Fisher linear discriminant analysis. *Department of Computer Science, University of Toronto*, 3(1).

West, D. M. (2013). *Air wars: Television advertising and social media in election campaigns*, 1952-2012. Sage.

Westmark, V. R. (2004). A definition for information system survivability. *In System Sciences, 2004. Proceedings of the 37th Annual Hawaii International Conference on* (pp. 10-pp). IEEE.

What is Big Data? and why is it important to me? (2015). *General Networks*. Available at http://www.gennet.com/big-data/big-data-important/

Wielki, J. (2013). Implementation of the Big Data concept in organizations-possibilities, impediments and challenges. *In Computer Science and Information Systems (FedCSIS), 2013 Federated Conference on* (pp. 985-989). IEEE.

Willis, O. (2016). Conservative media are already calling for Clinton's impeachment. Media Matters for America. Available at https://mediamatters.org/blog/2016/10/26/conservative-media-are-already-calling-clinton-s-impeachment/214133

Wladawsky-Berger, I. (2014). Why do we need data science when we've had statistics for centuries? Available at http://blogs.wsj.com/cio/2014/05/02/why-do-we-need-data-science-when-weve-had-statistics-for-centuries/

Wlodarczak, P., Soar, J., & Ally, M. (2015). What the future holds for Social Media data analysis. *International Journal of Computer, Information, Systems and Control Engineering*, 9(1), 16-19.

Xie, J., Yin, S., Ruan, X., Ding, Z., Tian, Y., Majors, J., & Qin, X. (2010). Improving MapReduce performance through data placement in heterogeneous Hadoop clusters. In Parallel & Distributed Processing, Workshops and Phd Forum (IPDPSW), 2010 IEEE International Symposium on (pp. 1-9). IEEE

Yamamoto, M., Kushin, M. J., & Dalisay, F. (2015). Social media and mobiles as political mobilization forces for young adults: Examining the moderating role of online political expression in political participation. *New Media & Society*, 17(6), 880-898.

Yeoman, J. (2017). Social MEDIA MARKETING FOR MUSICIANS: HOW TO GET FANS AND FOLLOWERS. Available at https://devumi.com/2017/01/social-media-marketing-for-musicians-get-fans-followers/

Yerbury, H. (2010), "Who to be? Generation X and Y in civil society online", youth studies Australia, Vol. 29 No. 2, pp. 25-32

York, A., (2016) How to successfully mine your social media data. Available at https://sproutsocial.com/insights/social-media-data/

Zaslavsky, A., Perera, C., & Georgakopoulos, D. (2013). Sensing as a service and Big Data. *arXiv preprint arXiv*:1301.0159.

Zhang, L., Stoffel, A., Behrisch, M., Mittelstadt, S., Schreck, T., Pompl, R., ... & Keim, D. (2012). Visual analytics for the Big Data era—A comparative review of state-of-the-art

commercial systems. *In Visual Analytics Science and Technology (VAST), 2012 IEEE Conference on* (pp. 173-182). IEEE.

Zhang, W., Johnson, T. J., Seltzer, T., & Bichard, S. (2010). The revolution will be networked: The influence of social networks on political attitudes and behaviors. *Social Science Computer Review*, 28, 75–92.

Zhou, L., & Wang, T. (2014). Social media: A new vehicle for city marketing in China. *Cities,* 37, 27-32.